AD-A249 990

DTIC
ELECTE
MAY 7 1992
S C D

1

IN-FLIGHT DECISION MAKING BY HIGH TIME AND LOW TIME PILOTS DURING INSTRUMENT OPERATIONS

BY

KENNETH LIVINGSTONE KEMPER

B.S., U.S. Air Force Academy, 1990

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1992

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

Urbana, Illinois

92-11978

92 5 01 011

Abstract

This study examined decision-making in high and low-time pilots (n=26) on a simulated IFR cross-country flight using MIDIS 3.0, a microcomputer-based flight-decision simulator. When confronted with situations which could endanger the safety and/or efficiency of the flight, it was hypothesized that high-time pilots would recognize the cues relevant to the problem, "pattern match" these cues with a situational schemata, or script, from long term memory (LTM), and choose to execute their first workable solution. It was hypothesized that low-time pilots would also attempt the same decision-making strategy, but because of their smaller experiential repertoires would fail to make a "pattern match" in LTM. It was posited that novices then are forced to use a "utility" strategy in which they must integrate cues with declarative knowledge, generate alternatives, evaluate outcomes, and finally choose the alternative calculated to bring the most utility. The difference in strategies was hypothesized to lead high-time pilots to choose more optimal solutions. The results clearly show that high-time pilots decision optimality is significantly better than low-time pilots. High-time pilots detected significantly more cues relevant to the problem, and chose the first alternative they considered significantly more often than their colleagues. The two groups were no different in basic information processing abilities, but were very different (with high-time pilots scoring significantly better) on tests measuring LTM based knowledge representations. Stepwise multiple regression analyses selected the pilot's certification level, **not total flight time**, as the best predictor of performance accounting for over half the variance. The LTM test measures were generally the next best predictors of performance.

ACKNOWLEDGEMENTS

This research was supported by a contract through the Idaho National Engineering Laboratory, EG&G Idaho, Inc., Idaho Falls, ID, through a subcontract with the Armstrong Aerospace Medical Research Laboratory, Wright Patterson Air Force Base, Ohio. We are especially grateful to Gary Reid, contract monitor, for his continuous support.

I would like to express my sincere gratitude to Dr. Alan F. Stokes for his guidance and consultation during this project. Without his advise and selfless commitment to this project, I would not have been able to finish within my given time constraint. In addition, sincere appreciation is extended to Dr. Janet Sniezek, my second reader, for her valuable input to this project.

I would also like to thank all of those individuals who made significant contributions throughout the course of this effort. Without Roger Marsh's programing expertise the development of the MIDIS 3.0 version and this experiment would not exist. Special thanks to Bert Henn and Tom Davis for their invaluable hours of assistance in developing the judges' rating scheme, and to Fred Kier for his countless hours assisting me in running subjects and quantifying data. I would also like to thank the CMI FAA Tower Supervisor and those personnel who served as expert judges for their generous support of this study. I am also grateful to Charles Hulin, Head of Pilot Training at the University of

Illinois Aviation Institute, and to those pilots who served as subjects in this study.

Finally, I want to extend my warmest appreciation to my family: Steve, Marybeth, Jason, Stephanie, Heidi, and Susan Kemper, and all of my wonderful friends. You kept my spirts up amidst these seemingly endless hours sitting in a little room behind a computer...Thank you.

Accession For
NTIS GRA&I ☑
DTIC TAB □
Unannounced □
Justification
By Per Form 50
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

QUALITY INSPECTED 4

TABLE OF CONTENTS

INTRODUCTION

Following Jensen and Benel's (1977) finding that 52% of all fatal "pilot error" accidents involved poor judgment, investigations of pilot judgment and decision making have increased within the aviation psychology community. However, much of the direct work on pilot judgment and decision making to date has been anecdotal, reconstructive, and based upon FAA accident and incident reports (e.g. Jensen & Benel, 1977; Jensen, 1982; Giffen & Rockwell, 1987), or intuition-based development of pilot judgment training programs (e.g. Buch & Diehl, 1984; Telfer, 1987; Connolly, Blackwell & Lester, 1987). The general consensus emerging from this body of literature is twofold: 1) Faulty pilot judgment is a contributing factor in the majority of "pilot error" accidents. 2) Empirical investigations of pilot judgment and decision making should be utilized in creating programs for judgment training and evaluation.

Despite these conclusions, researchers in this area generally have not attempted to bring a theory based empirical approach to bear on the issue of describing the processing activities that take place as decisions are carried out in the cockpit. This investigation will critically review the most relevant literature derived from anecdotal evidence, observational studies, and empirical/theoretical experimentation, and draw from this a rationale for the inquiry.

Anecdotal Observations

Everyday, in a variety of content area domains, countless numbers of complex, critical decisions are apparently made with incomplete or insufficient information and little time. Within the medical domain, for example, the skilled physician does not necessarily systematically compare each symptom with its likelihood of arising from a particular disease, but may instead perform a "pattern match" between a set of observed symptoms and the "syndrome" that is characteristic of a particular disease (Wickens, 1984). The syndrome is represented by a stored "prototype" of the disease in long term memory. Similarly, parole judgments (Carroll, 1980) and sentencing decisions (Ebbesen & Konecni, 1980) are described in much the same terminology.

Although the naive subject is likely to weigh exhaustively the implications of each independent fact of a specific case in order to arrive at a final judgment, expert performance is characterized by the use of extensive knowledge, organized as profiles or schemata. As in the medical example above, Carroll (1980) describes the parole judge's performance as a "pattern match" between the facts in a particular case and stored knowledge representations of the "alcohol abuser," the "aimless follower," the "heavy drug user," and other stereotypical profiles. Once evoked, a particular schema or profile then serves as a guide for future expectations, evaluations, and the acquisition of new information. Although a schemata-based decision strategy may expedite the decision process, it can harbor dangers of its own. For example, stereotypes sometimes lead to self-fulfilling prophecies and prejudices, which may bias the information-seeking processes needed to objectively evaluate a situation. The repercussions of this might include a parole judge's unjust sentencing, or a pilot jeopardizing the safety of a flight on the basis of a familiar but inappropriate schema.

Simmel and his colleagues suggest a two-stage process for the pilot's assessment of non-routine situations (Simmel, Cerkovnik, & McCarthy, 1987; Simmel & Shelton, 1987). Based upon anecdotal evidence from accident and incident reports, Simmel et al. hypothesize that upon encountering a non-routine event, the pilot bases decisions upon a diagnosis of the problem situation, followed by a determination of the potential consequences of that event. One of the main contentions of the two-stage theory is that most accidents and incidents are *not* due to inaccuracies in diagnosis. Rather, most are the result of the pilot's failure to assess accurately the potential consequences of the event. Accidents may result from an overassessment or underassessment of the seriousness of a given consequence.

Misassessment by the pilot presumably creates a strong situational stressor, often culminating in non-optimal decision making and deterioration of performance. One example from Simmel & Shelton (1987) illustrates this point. In this case, overassessment of the consequences of a nighttime alternator failure led the pilot to select a dangerous emergency landing at an unlit field (resulting in an accident), rather than landing at an alternative lighted airport only minutes away. Underassessment, on the other hand, can result in failure to take timely or appropriate action. Examples of the disastrous consequences of underassessment (for example, of the consequences of icing on the wing) are unfortunately all too familiar. Accurate assessment of consequences is affected by past experiences, which may provide "scripts" indicating what events are likely to follow a given situation. Scripts are event schemata or cognitive structures reflecting a predetermined, stereotyped sequence of actions that defines a well-known situation (Schank & Abelson, 1977). Each individual's particular experiences will influence a script's consequence assessments.

Simmel et al.'s two-stage theory seems intuitively satisfactory, but is not based on empirical research. There is a clear need to investigate, in a structured and empirical fashion, (i) how a pilot diagnoses a situation (e.g. what cues are used, what various options are considered plausible, etc...), and (ii) how he or she decides which viable option is best, and why.

Observational Research

Janis & Mann (1977) argue that defective decision making will result unless a full examination of all relevant options is completed, followed by the selection of the most optimal solution yielding the most positive outcome (i.e., the one which maximizes utility). In theory this would be ideal, but is likely to be impossible in settings with severe time restrictions. In contrast to normative decision-making models designed to optimize decision quality, Klein (1989) offers a Recognition-Primed Decision Model (RPD) to explain decision making under time pressure. He suggests that RPD is at the other end of the spectrum from the popular multiattribute utility analysis strategies, in that RPDs are non-optimizing, do not make comparisons among various options, and require little conscious deliberation.

The research done by Klein and his associates over the past five years has provided some insight into real-time, critical decision making. Klein has examined the strategies used by proficient decision makers under time pressure in natural settings (e.g. fireground command, military commanders battle planning, critical care nursing, corporate information management, and speed chess tournament play). On the basis of this work, Klein et al. argue that recognitional rather than analytical strategies predominate in decision making in operational contexts, (e.g., Calderwood, Klein, & Crandall, 1988; Crandall & Calderwood, 1989; Crandall & Klein, 1987; Klein et al., 1986; Klein & Thordsen, 1989; and Thordsen, Galushka, Klein, Young, & Brezovic, 1987). Recognitional strategies are characterized by the recognition of certain cues which prompt the decision maker to act upon the basis of familiar scenarios in long-term memory from past experiences. Analytical strategies, on the other hand, are characterized by integrating information in working memory, making inferences in order to predict the outcome of the situation. This involves the real-time generation of viable alternatives to best manage or thwart these outcomes, and the choosing of the alternative yielding the most positive effect, that is, the alternative with the maximum utility.

In urgent situations, there is evidence that this recognitional strategy may be used by expert decision makers. Klein, Calderwood, & Clinton-Cirocco (1986) found fireground commanders (FGCs) were less interested in an exhaustive review of the possibilities (and the selection of the very best option) than in finding an action that was "workable," "timely," and "cost-effective." Klein et al. estimated that more than 85% of the difficult, nonroutine, and high-risk decisions were made in less than a minute, and many only took a few seconds. Rarely did FGCs consider two options at once and choose the one with greater utility. Rather, the FGCs relied on their ability to recognize the cues associated with certain situations and to act on the basis of their past experience. They would use imagery to play out "scripts" (as defined by Schank and Abelson) until the **first workable solution** was found--and then they implemented it. If they had tried to generate many alternatives and systematically evaluate each one, the fires would have gotten out of control before a decision was ever reached. The strategy that Klein et al. observed is referred to as a "satisficing" strategy. Presumably, "veteran" FGCs would make decisions of higher quality in a shorter time than new FGCs, because of their larger repertoire of experiences.

Calderwood, Klein, & Crandall (1988) compared Master and novice chess players involved in chess games under no time stress (average move time 2.5 minutes) and under time-stressed, "blitz" games (average time per move only 6 seconds). The quality of the moves was rated by a chess Grandmaster. In this study, Masters' moves were not only significantly better than novices', but showed no decrement in blitz-time conditions. Novices, in contrast, showed drastic reduction in performance when under time pressure. Calderwood et al. concluded that more proficient players rely on recognitional decision making, which is less affected by time pressure than the analytical decision strategy adopted by novices.

In sum, the work of Klein et al. suggests that very different decision-making processes are used by experts and novices. Normative theory espouses that experts use a "utility-based" strategy to examine all possible options, and then choose the most optimal. Klein et al.'s observational research, however, supports a "satisficing" strategy, in which the first workable solution is chosen. They hypothesize that based on the cues of the situation, experts are able to recognize immediately the appropriate actions prompted by long-term memory. Novices, on the other hand, lack the wealth of "scripts" gained through past experience, and must integrate cues to analyze the situation.

Although these findings are interesting and important, they are nevertheless, based upon observational research rather than experimental evidence. As in the case of Simmel's work, experimental data is needed to clarify the status of this body of observational work.

Experimental Research

Definition of Expertise

An experimental approach to these issues was adopted in a series of experiments conducted at the University of Illinois' Aviation Research Laboratory. These studies adopt the convention of using the term "expert" to describe high-time experienced pilots, and the term "novice" to describe low-time pilots. Expertise proper is an elusive concept, however, and is not necessarily pinned down simply by the total time dimension. Wickens, Stokes, Barnett, & Davis (1987) criterion value for expertise was 400 hours merely because it divided the subject sample into approximately two equal groups.

Using a purely quantitative criterion, results in a loss of valuable information about the quality of previous flight experience. For example, one pilot may only have 350 total flight hours, but have an Airline Transport Pilot (ATP) certificate. Another could have 10,000 hours total time with only a private pilot license. Put another way "there is a difference between 100 hours of variegated experience and one hour of experience repeated 100 times" Stokes (1991). For this reason, subsequent studies (Barnett, 1989; Stokes et al., 1990) have collected biographical data in an attempt to tap these qualitative experiences. For example, pilots were asked to report the certificate held, total hours flown in actual instrument conditions, hours in single vs. multi-engine planes, and so forth. Experts, as defined by Barnett (1989), were those high-time pilots with a minimum of 1000 hours of flight experience with appreciable quality experiences. Conversely, novices were defined as low-time pilots with less than 500 total hours of flight experience.

Initial Experimental Research

Wickens, Stokes, Barnett, & Davis (1987) compared 38 instrument-rated pilots divided into two groups on the basis of flight hours. First, using paper and pencil tests, they assessed the putative information-processing components of decision-making ability (e.g., working memory capacity, logical reasoning, spatial ability, visual cue sampling skill, and declarative knowledge of instrument regulations and procedures). Stokes, Banich, Elledge, &

Ke (1988) converted these tests into an automated form known as SPARTANS (Simple Portable Aviation-Relevent Test-battery and Answer-scoring System) which has been used in all of the experiments outlined below. Wickens et al. (1987) attempted to predict pilot decision making by assessing non-domain-specific cognitive abilities in areas relevant to the general decision-making process. Subsequently, subjects flew an IFR (Instrument Flight Rules) flight on MIDIS, a Microcomputer-based Inflight Decision Simulator (Stokes, Wickens, & Davis, 1986; Stokes, 1991). Surprisingly, they found **no overall difference between the low-time and high-time groups** on the basis of decision performance. A few psychometric indices of decision making were moderately predictive of decision optimality in novices (working memory capacity, spatial abilities and declarative knowledge), but no indices predicted high-time pilot decision-making very well at all. The results are consistent with the view that alternative cognitive strategies may be employed by high-time and low-time pilots. Expert pilots may not need to rely upon working memory based skills as much as novices. The retrieval of domain-specific problem schemata or "scripts" from Long Term Memory (LTM) is presumably easier for "experts", due to their appreciable experiential repertoires, than to "novices". For example, consider a communications failure occurring in the clouds after being cleared for landing while on final approach. A high-time pilot might handle the situation quite quickly by recalling how he dealt with a similar problem before. A low-time pilot on the other hand, may not be able to draw upon this direct experience and may squander time evaluating the option to continue the approach or miss it immediately, while trying to recall the pertinent regulations and ground school instruction.

A possible explanation for the failure to find a difference in decision performance between low-time and high-time pilot performance is that subjects were asked to choose their decisions from up to six multiple choice alternatives presented to them on screen. Essentially, this design prompted the novices by providing "off the peg" hypotheses. By concentrating only on the output end of decision making, nothing was learned about differences in cue recognition or hypothesis generation. These deficiencies may account for the lack of differences between high-time and low-time groups.

Barnett (1989) focussed a study directly upon the possibility that experts and novices use different cognitive strategies. The specific hypothesis was that experts utilize direct retrieval of domain-specific problem "scripts" from long-term memory, whereas novices are more dependent upon working memory due to their lack of experience. This study compared 15 high-time and 15 low-time IFR pilots on the SPARTANS battery of domain-independent information processing tasks, three domain-specific tasks (Air Traffic Control (ATC) Recall Task, ATC Situation recognition, and Dynamic Diagnosis of Flight Problems Task), and a MIDIS flight. Once again, results showed that there were **no significant differences** in absolute performance level between the two groups. However, the three knowledge representation tests, designed to index representations of situational knowledge in long-term memory, were better predictors of high-time pilot performance than the information-processing battery tests. Additionally, the results found information processing measures did no better in predicting performance of low-time pilots than of the high-time pilots: decision performance was not strongly predicted by the information processing tests. It is an important fact that, domain-specific tests were the best predictors of flight decision making for both groups.

Again, the peculiar homogeneity between the two cohort groups may be due to the "prompted" decision format. Additionally, the 2-strategy theory was not strongly supported. The finding that LTM based measures worked best for both experts and novices might suggest that both groups attempt to use a "satisficing" strategy when possible, and only revert back to working memory strategies when they have to. Presumably, under this hypothesis, experts would make better decisions in a more timely manner due to their vast experiential repertoire of situations. Furthermore, the domain-independent (SPARTANS) tests may yet be predictive of novice performance if the novices were made responsible for spotting cues and then generating their own action alternatives. Under this format, the analytic working memory based strategy would presumably not be bypassed by supplied hypotheses.

Pilot Decision Making Under Stress

The most recent study in this series (Stokes, Belger, & Zhang, 1990) contrasted "novice" and "expert" instrument pilots to test whether alternative cognitive strategies are used by "expert" vs "novice" pilots in stressful and non-stressful situations. It was found that the two groups' degradation of performance under stress was roughly equivalent and on the same cognitive abilities **for the domain-independent measures of information processing.** However, only novices exhibited performance decrements in the domain-specific (i.e., operational flight) task under stress, and only on dynamic scenarios involving attention to moving display indicators. Since trait anxiety tests showed equal mean scores for the expert and novice groups, the difference in performance on the operational task under stress was apparently not due to a personality variable, an inherent "stress resistance" possessed by experts. Rather it supports the hypothesis that novice performance is more dependent than expert performance on cognitive operations in working memory (which is degraded under stress). The experts presumably had less of a problem because they were more able to draw on LTM based situational schemata to diagnose and react to situations, without having to resort to stress prone analytical processes in working memory.

Rationale for this Study

The anecdotal, observational, and experimental investigations reviewed here provide converging evidence that experts make decisions in a qualitatively different manner than novices. However, there are areas of disagreement or contradiction. Observations from Klein et al., for example, suggest that experts make better decisions in less time as compared to novices (Calderwood, Klein, & Crandall, 1988; Crandall & Calderwood, 1989; Crandall & Klein, 1987; Klein et al., 1986; Klein & Thordsen, 1989; and Thordsen, Galushka, Klein, Young, & Brezovic, 1987). However, these findings have not been confirmed in the experimental, empirically based research of Wickens, Stokes, Barnett, & Davis, 1987, Barnett, 1989, and Stokes, Belger, & Zhang, 1990. It is hypothesized here that the multiple choice format of the experimental research may have attenuated variance, or may have prompted novices to search for cues they otherwise may have overlooked in the decision process. If so, this would have negated the expert's "script-based" advantage. Obviously, in an actual emergency during an instrument flight there is no computer to advise that something is wrong and offer two to six alternatives from which to chose. Additionally, the initial experimental research (Wickens et al., 1987) assumed experts employ utility strategies when decision making. They may have simply picked the option closest to the first practical solution they thought of without comparing outcomes of all options.

Stokes' most recent effort (1991) offers a model describing the in-flight decision making process. Figure 1 illustrates the hypothesized path pilots follow during in-flight decision making. The model posits that pilots' initially attempt to use LTM to match problem cues with similar situations experienced in the past. If a pattern match is not made, the pilot is forced to use working memory to integrate cues, and to generate and evaluate viable action alternatives. The model is consistent with the results of studies detailed in the previous literature review, but, additional empirical evidence is needed to substantiate it.

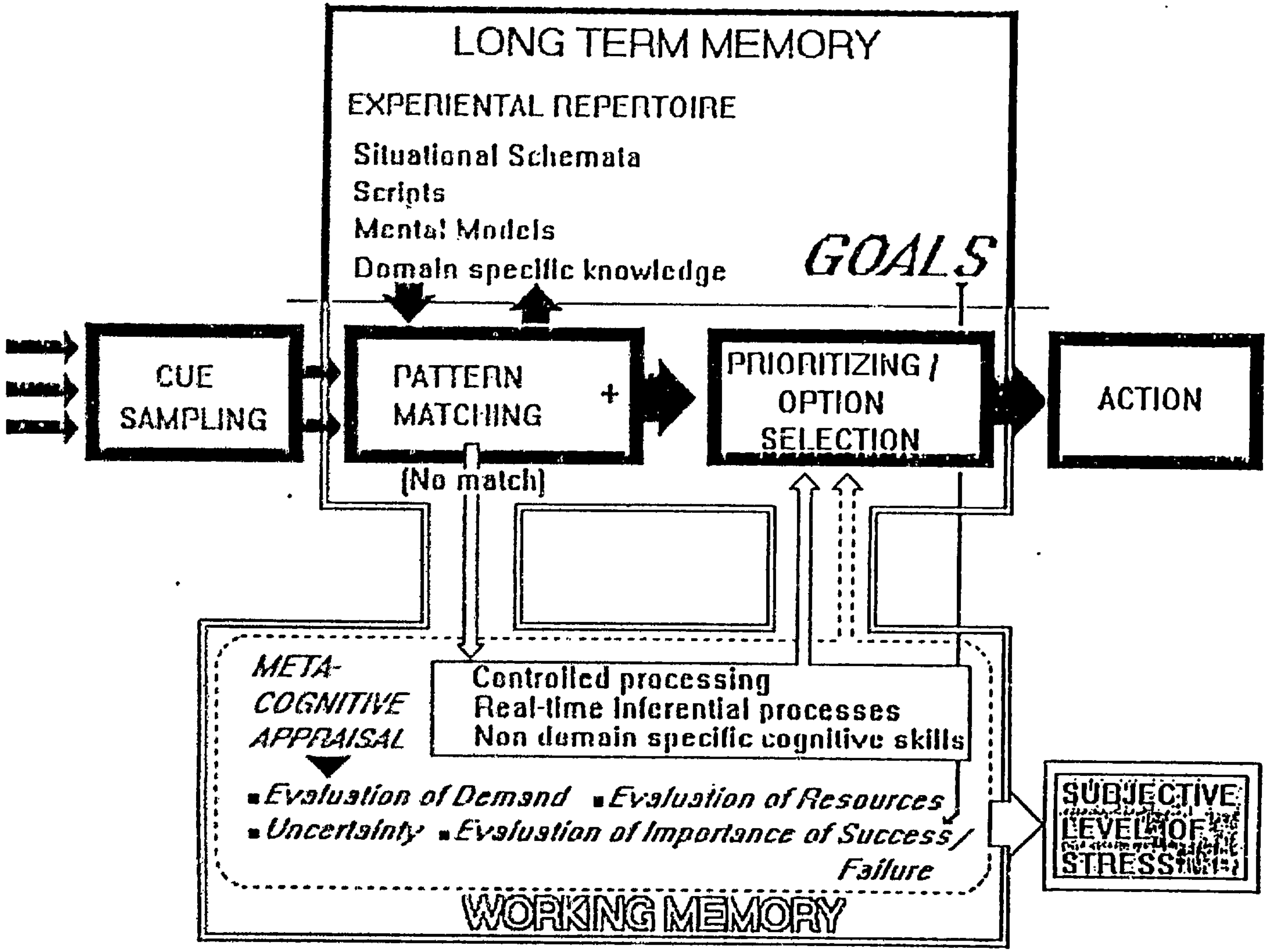


Figure 1. Stokes' Pilot Decision Making Model

The first goal of the present study was to overcome certain specific methodological limitations the earlier work. This was accomplished using MIDIS 3.0, a much modified version of the flight simulation system which allows "open ended" responses to be keyed in by pilots. Essentially, this eliminates the prompts to cue recognition and forces the pilots to generate their own alternatives for the situation (Stokes, 1990; 1991). Second, the study was designed to allow observation of decision strategies that do not fit the "utility theory" model.

Hypotheses

The central inquiry of the investigation is to discover how high-time pilots differ from low-time pilots with respect to cue recognition, hypothesis generation, option selection, and optimality of choice. It is posited that high-time pilots use the "satisficing" strategy associated with LTM more often, while low-time pilots will be forced to use a utility strategy associated more with working memory. Nine main hypotheses will be tested:

i. There will be few or no overall differences between high-time and low-time pilots' information processing abilities.

ii. Domain specific (LTM based) tests will be better overall predictors of pilot decision making performance than non-domain specific information processing tasks.

iii. There will be significant differences between the two groups on the long term memory representations measures. Specifically, high-time pilots are hypothesized to perform better on the ATC Recall Task and the ATC situation recognition Task.

iv. In contrast to Simmel et al.'s theory (1987), the source of inferior decision making in low-time pilots lies in poorer cue recognition. Therefore low-time pilots will report fewer correct cues and more irrelevant cues when analyzing a situation than high-time pilots;

v. The satisficing strategy of high-time pilots will result in low-time pilots generating more alternatives than high-time pilots, because high-time pilots will use their first workable alternative and low-time pilots will consider the utility of many alternatives;

vi. The satisficing strategy will lead high-time pilots to choose the first alternative generated as the most optimal solution more often than low-time pilots;

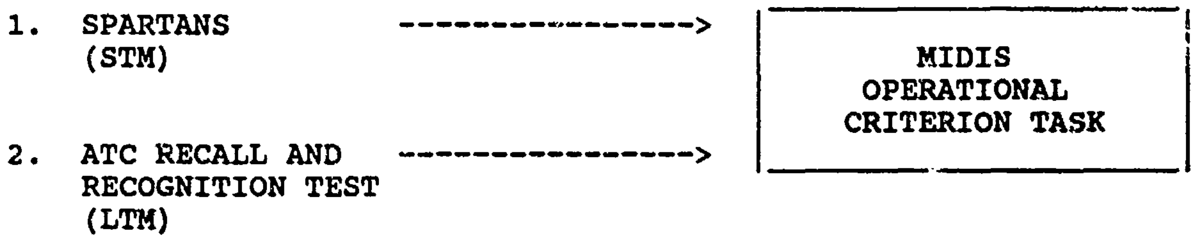
vii. Because low-time pilots are forced to use a Short Term Memory (STM) based strategy more often, their performance (in quantity and quality of cue recognition and alternative generation) will be predicted by the general cognitive abilities pretests, while high-time pilots' performance will not;

viii. Because high-time pilots are hypothesized to use a LTM based strategy, their performance will be predicted best by schemata based measures rather than domain independent cognitive tests.

IX. High-time pilots' decision optimality will be greater than low-time pilots' once the prompting element of the decision task is removed.

METHOD

The general approach was first to measure short term memory (STM) information processing abilities and long term memory (LTM) knowledge based representations of situations in both high-time and low-time pilots. The next step involved determining the predictive power of each test on the pilots' performance in a simulated flight in instrument conditions. The SPARTANS battery of information processing tests described earlier was used to measure STM based information processing abilities. Recall of LTM based domain specific knowledge and situational schemata was measured using tape recorded ATC messages. The ATC recall task required subjects to listen to ATC messages, then write as much of the message as they could remember down on paper. The ATC situation recognition task required subjects to choose diagrams representing situations heard on the radio between ATC and pilots. The criterion task, the MIDIS flight, was administered via a desktop computer simulation. An outline of this methodological approach is presented schematically below.



1. SPARTANS (STM) ----------------->

2. ATC RECALL AND RECOGNITION TEST (LTM) ----------------->

MIDIS OPERATIONAL CRITERION TASK

Cognitive Abilities Batteries

The cognitive performance batteries were administered to determine whether the two pilot groups were fundamentally equal on non-domain specific cognitive abilities (STM) and domain-specific abilities (LTM). The first LTM test is a verbal assessment entitled the ATC Recall Task. This task involves reconstructing both randomized and coherent radio call sequences from memory. Presumably, the quality of reconstruction is primarily influenced by the availability of appropriate situational "scripts"--memory effects are controlled out using recall of the jumbled transmissions. The second LTM task attempts to tap one's spatial or mental picture of a situation. The ATC Situation recognition Task involves listening to various ATC calls, building a mental picture of the situation, and selecting the appropriate diagram of the scenario.

Information Processing Measures (STM)

Sixteen sub-tasks relevant to pilot decision making (SPARTANS) were used to index individual differences in the efficiency of short-term processes in working memory. The compiled test battery consists of a one-to-one mapping between cognitive attributes relevant to pilot judgment and cognitive tests designed to measure each individual attribute. Most sub-tasks were derived from the Educational Testing Service (ETS) kit of Factor-Referenced Cognitive Tests. Others were developed within the University of Illinois' Aviation Research Laboratory (Wickens, Braune, Stokes, & Strayer, 1985; Stokes, Banich, & Elledge, 1988;). A brief description of each task follows.

1. **Spatial Memory Task.** In this test the subject views an inspection set of nonsense figures. These are abstract amoeboid figures without geometrical or pictorial significance that would facilitate verbal recording. About twenty minutes after initial presentation subjects view 40 figures and decide for each one whether or not it had been a member of the inspection set viewed previously (Banich, Stokes, & Karol, 1990).

2. **Hidden Patterns Recognition.** This task was adapted from a test presented in the ETS (Education Testing Service) Manual for the Kit of Factor-referenced Cognitive Tests (Eksttrom, French, & Harmen, 1976), this task assesses flexibility of closure and factor loads with spatial ability. Subjects must detect an abstract line drawing embedded within a more complex pattern of lines.

3. **Rotation Hidden Patterns.** Identical to the previous sub-task save that the target figure may be rotated. Mirror images, however, are considered non-targets.

4. **Maze Tracking Task.** This spatial task adapted from ETS presents subjects with a series of line mazes of increasing complexity. Each maze must be cognitively traced as rapidly as possible to decide whether or not there is an unbroken path from beginning to end. As in most of the sub-tasks in the battery, both accuracy and latency data are recorded.

5. **Sternberg task.** The Sternberg Task (Sternberg, 1966) is a standard memory search task in which a number of target letters are memorized and a series of stimulus letters are presented. Subjects must depress one of two joystick buttons to indicate whether or not a stimulus letter is a member of a previously presented target set. The memory set size used in SPARTANS is four.

6. **Zero Order (Position) Tracking Task.** This is a psychomotor test requiring subjects to keep a vertical sinusoidal line between two horizontally movable cursors using the joystick. This task is an adaptation of that developed in the Netherlands by Boer and Gaillard (1986).

7. **Dual Task Zero Order Tracking.** This task combines the Sternberg task with the zero order tracking task and enables performance decrements due to dual-task timesharing to be observed.

8. **First Order (Velocity) Tracking Task.** This task is similar to, but more difficult than the zero order version. The cursor moves horizontally at a constant rate with inputs from the joystick. Movement to start and stop the change in direction are necessary.

9. **Dual Task First Order Tracking.** This task combines the Sternberg task with the first order tracking task.

10. **Risk-Taking Task.** The box task assesses predisposition to risk-taking. Subjects may select up to ten boxes containing hidden scores. The score, once revealed, is added to the subject's box-task total. The subject knows, however, that one random box per trial is "booby-trapped". If selected, the subject's entire score for the trial becomes zero. Over a 20 trial run the mean number of boxes selected as well as a measure of risk taking consistency is measured.

11. **Stroop Task.** This classic test assesses selective attention conceptualized as the ability to disregard or tune out irrelevant information (Stroop, 1935). In this implementation of the test, words denoting color terms ("blue", "red", "yellow", etc.) are displayed singly on the screen. In a control condition the color terms are displayed in black letters. In the "interference" condition, the words are displayed in four, randomly selected, primary colors that conflict with the color that word denotes. Using a joystick operated pointer, the subject must match the color denoted by a term (but not its display color) with one of four color patches displayed at the bottom of the screen. The amount of interference is indexed primarily by the difference in time between responses in the control and the experimental condition. Other variables measured are latency in monochrome and color responses, and correct response decrement.

12. **Nonsense Syllogism.** This logical reasoning task was adapted from the ETS test. The subject is presented with a series of

syllogisms and must decide in each case whether the conclusion is valid or invalid.

13. **Visual Number Scan.** This task indexes working memory capacity. Subjects are presented with a digit span task in which a string of digits are presented visually. String length increases as trials progress. After the screen is cleared the subject must type the list back into the computer.

14. **Backwards Visual Number Scan.** This task is the same as the visual number scan described above with the exception of entering the string of digits back into the computer in reverse order. It has been maintained that while number span forward is a left hemisphere task, number span in reverse is a right hemisphere spatial imagery task (Banich, Stokes, & Elledge, 1988).

Knowledge Representations Measures (LTM)

Three tests were employed to assess individual differences in availability of domain specific situational schemata or knowledge representation. First a declarative or "textbook" knowledge measure, the "FAA Quiz", was administered to both high and low-time pilots. The test consisted of 25 questions from the FAA Instrument Written Test Manual covering a balanced range of topics (weather, chart use, systems, etc...), but excluding procedural knowledge questions.

Procedural knowledge was evaluated using the method developed and used by Barnett (1989) and Stokes et al. (1990). These researchers administered two domain-specific tasks intended to index representations of situational knowledge in long-term memory. The tasks were devised as direct analogues of the well-known chess experiments reported by de Groot (1965), and Chase and Simon (1973). These experiments showed that chess Masters were far more capable than novices of accurately replacing chess pieces on a board **if those pieces had been arranged in a coherent game position.** If the pieces had originally been randomly placed, chess novices and Masters were equally poor at replacing them. This result has been ascribed to the large repertoire of game states that chess Masters have internalized.

Recall Task

As Barnett (1989) and Stokes (1990) have described, the Air Traffic Control (ATC) Recall Task used radio calls in place of chess pieces. Subjects were presented with a series of ATC messages and instructions (e.g. taxiing instructions, approach and landing instructions, or clearance information). For each trial, the subjects were given as much time as they needed to familiarize themselves with a diagram of an airport or approach plate, as well as directions indicating their present position. Subjects then listened to a tape-recorded message relating the calls. Immediately following the message, subjects turned the diagram over so they could no longer see it and recalled from memory as much of the exact radio call sequence in the correct sequence as they possibly could. Twelve radio calls were presented in all (averaging six lines of text per trial).

In a fashion similar to the Chase and Simon method, half of the radio call sequences followed a coherent sequence concerning takeoffs, approaches, or navigation, while the other half consisted of arbitrary randomized calls selected from a variety of different flight scenarios. Simple memory effects were controlled by finding the difference between the total number of items recalled in the sequential presentation and the number of items recalled from the random instructions. Examples of a coherent call and a randomized call (along with corresponding diagrams) are found in Appendix A.

Situation Recognition Task

The ATC Situation recognition Task was used to further differentiate between pilots able to construct an accurate mental representation of operations within the airspace from those who cannot. This task required subjects to listen to ten radio exchanges between ATC and other pilots. Following the tape-recorded sequences, subjects opened a booklet with four diagrams for each scenario and selected the appropriate diagram which best depicted the situation giving rise to the exchange of calls. Each diagram placed the various aircraft involved in the exchange in different positions--only one of the four illustrations was correct.

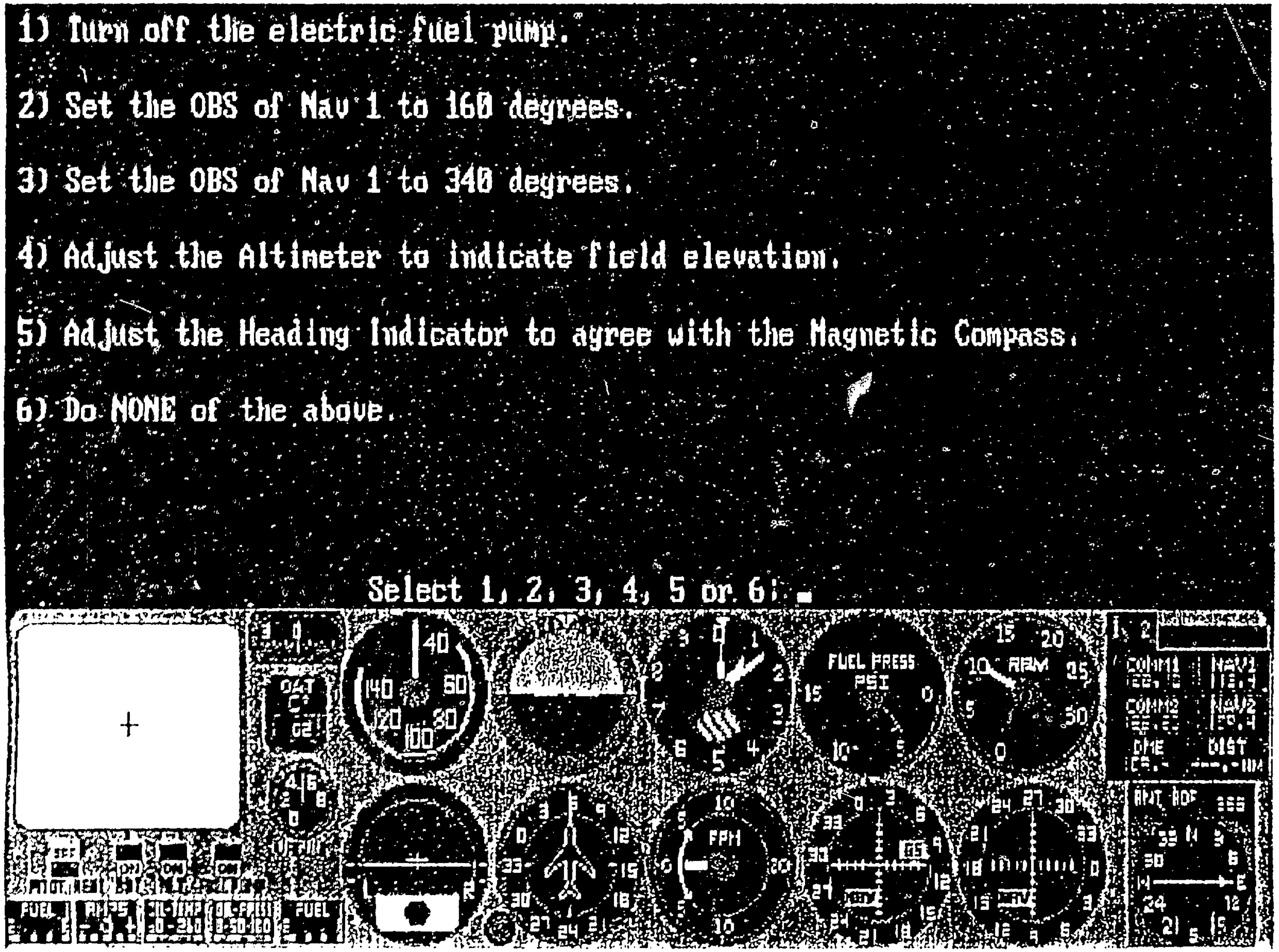
The MIDIS 3.0 System: The Task

Version 3.0 of the MIDIS simulation (Stokes, 1990; Stokes, 1991) has a full, high-fidelity instrument panel based on a Beech Sport 180, the type of aircraft used for training at the University of Illinois Institute of Aviation. This display, implemented via the HALO graphics package and 16 color Enhanced Graphics Adapter, represents a full IFR panel with operating attitude, navigational, and engine instruments. The MIDIS software allows the readings on the instrument panel to change throughout the course of the "flight" in synchrony with the prevailing scenario. These changes may occur either discretely or continuously. MIDIS does not attempt to simulate the flight dynamics of an aircraft from control inputs. Rather it imposes judgment requirements by presenting a series of time slices or "scenarios" in the course of a coherent unfolding flight. The present version of MIDIS is designed to follow a linear path on an IFR flight from Madison, Wisconsin to Mason City, Iowa through a pre-set sequence of scenarios, although it has been so constructed as to appear to be an open ended flight in which the pilot controls the sequence.

A scenario can be defined by either the instrument panel together with a text description of particular circumstances, or by a particular normal or abnormal configuration of the instrument panel alone. These two representations are known as static and dynamic scenarios, respectively. Where text accompanies the panel, the instruments are stable (i.e., they show no rate of change). Figure 2 illustrates a static sample MIDIS screen.

In dynamic scenarios, where there is no text, the instruments can show a rate of change. This allows us to study an important class of decisions--those involving the detection of changes and the integration of decision cues in real time. Any scenario may represent a problem or it may not. A problem scenario is one in which the circumstances have clear and present implications for the efficiency or safety of the flight, requiring diagnostic and corrective action to be taken. For example, it may involve a navigational problem, a meteorological problem, a systems problem or an operational problem.

Figure 2. Sample MIDIS Screen -- Multiple Choice Problem Senerio



1) Turn off the electric fuel pump.

2) Set the OBS of Nav 1 to 160 degrees.

3) Set the OBS of Nav 1 to 340 degrees.

4) Adjust the Altimeter to indicate field elevation.

5) Adjust the Heading Indicator to agree with the Magnetic Compass.

6) Do NONE of the above.

Select 1, 2, 3, 4, 5 or 6:

After viewing the static display describing the scenario, subjects pressed the return key to request the decision options. Subjects select a numerical option with a keypress, followed by a second numerical keypress to indicate confidence on a scale ranging from 1 to 5. Subjects were given instructions to rate the confidence in their decision based on the following scale:

5 -- Absolutely certain
4 -- Reasonably sure
3 -- Some room for doubt
2 -- Rather uncertain
1 -- Little more than a guess

Subjects were encouraged to use the full scale as appropriate. The confidence response automatically steps the program forward to the next flight scenario (which may or may not be contingent upon the nature of the confidence response).

When a dynamic scenario is viewed (e.g., portraying steady state flight through turbulence, or recovery from an evasive maneuver), subjects are allowed to press a red key to indicate whether they believe that an abnormality had occurred. After the dynamic scenario is played out (usually 1-3 minutes), assuming that a failure actually has occurred, an "open ended" response format appears. The open ended scenarios require subjects to key in what cues they used that tipped them off that there was a problem. They may then type, for example, "RPM decreased <return>, fuel pressure decreased <return>" and then press a key to go to the next screen. This screen requires subjects to list all plausible action alternatives, then they are asked to identify the best of these alternatives, and finally enter their confidence in the decision form 1-5.

Altogether, fifty-one scenarios were presented in the experimental criterion flight, 25 of these were dynamic, 8 were static, and 24 required open-ended responses. In addition to those dynamic scenarios that involved a problem, the flight consisted of a number of episodes of non-problem flight, preserving some of the natural dynamic characteristics of normal flight. A timer indicating the number of seconds spent on each problem coincided with a little beep to serve as a mild time pressure instrument.

Seven other performance variables are monitored, most of them unobtrusively. Four of these relate to response selection: decision choice, optimality, decision time (latency), and decision confidence. While subjects are reading the program instructions a mean reading speed is calculated as the number of syllables read per second. Scenarios and options are then analyzed for word and syllable counts. In this manner, individual differences in reading speed can be factored out of the latency data.

Optimality Coding

The multiple choice MIDIS scenarios were coded by a team of flight instructors as to the optimality of each response on a scale from 1-5 (Stokes et al., 1987). The best option was assigned a value of 5. The other alternative received values ranging from 1-4, depending upon how close they were to being plausible alternatives.

The "open-ended" responses required four steps. First the pilots listed all problem cues they noticed for that particular scenario. Then they listed all the plausible action alternatives appropriate when in that situation. Third, they were asked to identify what they considered the best alternative on their list. They also provided a rating from 1-5 upon their level of confidence in this choice. And finally, they were asked to give a brief rationale explaining why that alternative was chosen.

Each problem scenario was designed with one to five critical problem cues that subjects must identify to fully diagnose the situation. For example, the flight incorporates a scenario in which the aircraft is climbing through the known freezing level with the outside air temperature below freezing and structural ice is developing. To diagnose this, the pilot needs to recognize that (1) the airspeed apparently decreased, (2) the attitude indicator was showing a pitch up while the aircraft continued level at the correct altitude, and (3) RPM had decreased. Only noticing one or two of these cues is likely to cause misdiagnosis. Therefore, each open-ended problem was first analyzed for correctly identifying a problem cue, missing one or more of the problem cues, and erroneously identifying an irrelevant cue.

Response Judging

Next, a panel of three judges, acting independently, rated the optimality of each alternative based on its effect upon the safety and/or efficiency of the flight. Two of the judges were very experienced high-time instructors from the University of Illinois' Aviation Institute, while the third was a professional pilot who was also an experienced Air Traffic Controller from the Champaign tower). None of the judges had acted as subjects. The judges were provided with all the pertinent information for each scenario giving them a "God's eye view" of the situation. That is, they knew exactly what the situation was, what failures had in fact occurred and what corrective actions were or were not appropriate. The judges then rated the alternatives listed based on safety and/or efficiency using a scale ranging from 1-7; 7 being the best possible action in terms of safety and/or efficiency for the given situation. In the structural icing example, the best alternative, given all known information, is to request amended routing and to decend to a lower (and warmer) altitude. A less optimal solution for this scenario would be to attempt to climb the iced up aircraft in an effort to get above the cloud tops. The ratings were defined for the judges using the scale below:

7 -- The best possible action that could be taken in this situation
6 -- A good, plausible action which will increase safety and/or efficiency
5 -- An acceptable action which will increase safety and/or efficiency
4 -- A neutral action; one which will neither increase nor decrease the safety and/or efficiency of the flight
3 -- A poor action which may be detrimental to the safety and/or efficiency of flight
2 -- A bad action which will be detrimental to the safety and/or efficiency of flight
1 -- A very adverse action which will seriously endanger flight

Procedure

Data collection was carried out in two sessions for each subject. In the first session, lasting approximately two hours, a condensed SPARTANS battery of cognitive abilities and two LTM tests described previously were administered. In the second session, conducted one to five days later, subjects completed the flight task on the MIDIS simulator.

During this second session, subjects were instructed to plan an IFR flight from Madison, Wisconsin to Mason City, Iowa with a normal balance of safety and efficiency. Sectional charts, L-charts, approach plates, pilot operating handbooks, and a flight service station weather briefing were provided for flight planning. Although no "stick-and-rudder" flying is involved in the MIDIS simulation, the pilot must be familiar with various performance characteristics of the aircraft in order to flight plan and make accurate decisions throughout the flight. Therefore, those pilots who were unfamiliar with the aircraft simulated by MIDIS, the Beech Sport 180, were briefed on the performance characteristics of the aircraft and given an opportunity to review a diagram on the instrument panel before flight planning. The subjects were given a time limit of 30 minutes in which to plan the flight.

Following the flight planning, pilots undertook the simulated flight on the MIDIS system. MIDIS initially provides subjects with an instructional overview, which is followed by a system familiarization flight and finally the actual flight. Data collected for each judgment scenario include an optimality rating of the decision option selected, decision latency and the pilot's confidence rating for each judgment. The total time for the second experimental session (including flight planning) was approximately 2.5 hours.

Subjects

Thirteen student pilots presently enrolled in classes working toward an instrument certificate served as the low-time group, while eleven University of Illinois Aviation Institute instructor pilots, one corporate, and one military pilot were selected from a pool of 30 applicants to serve as subjects in the high-time group. All subjects brought their pilot log books and filled out a biographical information sheet found in Appendix B to determine flight experience. Rather than classifying subjects into two groups on a post hoc basis, subjects had to qualify for each group. High-time pilots had to have a minimum of 1,500 hours of flight experience and hold at least flight instructor, commercial, and instrument certificates. To qualify for the low-time group, the pilots had to have a private pilot license and either be working on an instrument certificate, or have an instrument certificate with less than 50 hours total instrument time.

The low-time pilots' total hours ranged from 65 to 150 hours with mean total hours = 105.77, mean instrument hours (including simulated) = 31.0, and mean actual instrument hours = 2.77 hours. High-time pilots were very experienced coming from a variety of backgrounds including commercial airlines and the U.S. Air Force. Nine had Airline Transport Pilot (ATP) certificates, while the other four had at least flight instructor, commercial, and instrument certificates. Total time ranged from 1,710-10,800 hours with mean total hours = 4,940, mean instrument hours (including simulated) = 772.67, and mean actual instrument hours = 322.17. Table 1 shows ranges, means, and total hours for various conditions and aircraft.

Table 1.

BIOGRAPHICAL FLIGHT INFORMATION
(REPORTED IN HOURS)

High-time pilots	**Low-time pilots**
Certificates:	
9 Airline Transport Pilot 4 CFI, Instrument, and Commercial	3 Instrument certificate 9 Private pilot license
Average Total Time (Range):	
4940 (1710 to 10800)	105.77 (65 to 150)
Instrument Hours (Range):	
772.67 (190 to 2000)	31.00 (10 to 75)
Last 90 Days (Range):	
VFR 67.67 (5 to 150) IFR 37.33 (2 to 300)	VFR 9.5 (1 to 15) IFR 12.5 (3 to 20)
Total Time in Actual Instrument conditions (Range):	
322.17 (108 to 929)	2.77 (1 to 9)
Light Single Engine (Range):	
3312.25 (1295 to 9859)	105.77 (83 to 150)
Light Multi Engine (Range):	
658.25 (10 to 1200)	0.0
Heavy Multi Engine (Range):	
726.58 (0 to 5000)	0.0

RESULTS

Initial analyses will compare the high-time against the low-time pilots to identify global differences in ability profiles or MIDIS performance. Following these primary inquiries, the discussion will turn to an exploration of patterns of performance predictors within each group.

Between-Group Analysis of Test Battery Scores

Based on past experiments (Wickens et al., 1988; Barnett, 1989; Stokes et al., 1990), few differences were expected between the two groups on the information processing measures. In contrast to the findings of Barnett (1989), (i.e., no significant differences exist between high and low-time groups on knowledge representation measures, it was hypothesized that the two groups would differ as found by Stokes et al. (1990), who used an improved and extended ATC test. Stokes et al. found that (in the ATC Recall task) experts showed a much larger improvement in scores from random to coherent conditions than novices .

Information Processing Abilities

Analysis of SPARTANS test battery scores for the two cohorts were computed using a two-sample t-test. Age was expected to be a confound due to a mean age of 20.62 for low-time pilots, and mean age of 39.7 for the high-time pilots. As suspected, initial analyses found significant differences between the two cohorts on response latencies. Therefore, an analysis of covariance was conducted factoring out Age. The results were then consistent with the hypothesis that few, or none, of the information processing measures would differ significantly for the two groups. In fact, the two cohorts failed to differ on any of the cognitive information processing measures. In light of this information, it can be assumed that the two cohorts are essentially equal in terms of information processing ability. Hence, any group differences in criterion performance cannot be attributed to age or an inherent ability difference possessed by one of the groups.

Domain Dependent (LTM) Knowledge Representation Measures

As found in previous studies (Stokes et al., 1990; Barnett, 1989), there was no difference between high-time and low-time pilots on declarative knowledge of instrument flight regulations and procedures as measured by standard FAA questions. As hypothesized however, significant differences on knowledge representation measures were found between the two groups on the ATC Situation Recognition and ATC Recall tasks.

ATC Situation Recognition Results

Analysis of the ATC Situation Recognition task revealed a difference between groups with high-time pilots scoring significantly better than low-time pilots ($t(24) = 3.47$, $p < .002$). Table 2 lists the observed ATC recognition scores by the number correct out of ten.

This difference in the hypothesized direction supports the notion that high-time pilots are better able to draw on experiential repertoires via situational schemata. An argument could be made that the ATC Situation Recognition task was, in fact, merely reflecting spatial ability, a STM based skill, and is not a true reflection of any long term memory mechanism or process. This is not the case, however. First, a correlational analysis pairing

ATC Situation Recognition score with each spatial variable in SPARTANS found **no significant relationships.** Second, evidence suggests the ATC Situation Recognition task is truly tapping ability to utilize domain-specific knowledge in LTM. Results from multiple regression analyses, discussed later, mark the orthogonal nature of the ATC Situation Recognition task and the Spatial variables.

Table 2.

ATC Situation Recognition Task - Number Correct

High-time pilots	7.77
	[t (24) = 3.47, $p < .002$]
Low-time pilots	5.85

ATC Recall Results

The ATC Recall task analyzed the difference score between the number of concepts and words recalled in the coherent and random message conditions (Table 3 lists the means for each group and associated p-values). This permitted the effects of individual working memory capacity, hearing impairment, etc., to be controlled out. A significant effect for number of concepts recalled was found, however, no such difference was found for the word analysis due to noise created by function words: of, and, with, etc.

The number of words recalled verbatim from the messages was not significant, although high-time pilots did recall more. This measure was flawed by giving equal weight to each word. Low-time pilots appeared to recall much of the message. Upon closer examination, however, many of the words were merely functional prepositions. The high-time pilots, on the other hand, tended to leave out function words and remember only key phrases.

Table 3.

ATC Recall Task - Concepts and (Words) Recalled Difference Scores

	Concepts	Words
High-time pilots	1.29	(3.25)
	[t (24) = 2.89, p<.008]	[t (24) = .27, p<.893]
Low-time pilots	.58	(3.01)

The difference found in number of concepts recalled is consistent with long term memory based hypotheses. High-time pilots were hypothesized to recall significantly more concepts when presented with coherent ATC messages versus random ATC messages. High-time pilots did, in fact, have a recall score difference .75 greater than the low-time groups. This finding supports the reasoning that low-time pilots are not able utilize templates or scripts for various types of situations. High-time pilots, on the other hand, were able to realize that it was "one of those type of messages" and simply fill the new information into a generic template.

Between-Group Comparison of MIDIS Scores

Performance on the MIDIS flight was analyzed for both the multiple choice and open-ended decision formats to determine whether high-time and low-time pilots differed significantly from one another in terms of these dependent variables. For the first time in the series of MIDIS experiments conducted since 1986 at the

University of Illinois, significant differences were revealed between high-time and low-time pilots decision optimalities. An investigation follows to determine why the difference between the two groups existed.

Multiple Choice Decision Optimality

Decision optimality for the multiple choice questions was computed by averaging the optimality associated with each of the nine questions. High-time pilots performed significantly better than the low-time pilots on the mean optimality of multiple choice responses. Out of a possible maximum of 5.0, high-time pilot's average multiple choice optimality was 3.485 versus 2.999 for the low-time group [t (24) = 2.39; p<.25].

Decision Optimality of Open-Ended Responses

Inter-rater reliabilities of the independent, "blind" judges were calculated to determine the credibility of the results. Reliabilities were quite high considering the subjectiveness of the task. All correlations were very significant (p < .001) between the judges. Table 4 lists the reliabilities and associated p-value significance level of each correlation. Given the high inter-rater reliabilities, it is reasonable to combine the judges into an overall optimality rating. However, they were kept separate for the following analyses to act as a "built-in" replication.

Table 4.

Inter-rater Reliabilities: Open-Ended Response Optimalities

Judge 1 vs Judge 2	r=.9223	p < .001
Judge 1 vs Judge 3	r=.8161	p < .001
Judge 2 vs Judge 3	r=.8365	p < .001

Group Optimalities

It was hypothesized that high-time pilots' decision optimality would be greater than low-time pilots' once the prompting element of the decision task was removed. High-time pilots did, in fact, perform significantly better, in terms of safety and efficiency of open-ended alternative selection, than low-time pilots. This finding was very significant (p < .001) for all three independent judges. Table 5 lists each judge’s optimality percentage and associated p-values.

Table 5.

MIDIS Performance -- Open-Ended Response Optimalities

	High-time Pilots	Low-time Pilots	Significance
Judge 1	.7285	.6100	t (24) = 6.15; p < .001
Judge 2	.6877	.5485	t (24) = 5.91; p < .001
Judge 3	.7677	.6431	t (24) = 5.00; p < .001

The judges’ scores are given as a percentage out of the total possible score. For example, if a subject received a seven on every scenario then they would have gotten all 153 points possible and received a hundred percent. The scores were summed for each alternative and divided by 153 to reach the subject's percentage. If the subject took no action by failing to offer any alternatives, then a score was given which reflected the severity of not taking any action. In some cases, failing to take action was given a neutral rating, while in others taking action was critical to the safety of the flight, and therefore no action was given a lower rating. At no time was failing to take an action given a positive rating (i.e., a five, six or seven). This created a floor effect for each judge. For instance, if the subject never listed an alternative the entire flight, their optimality for judges 1, 2, and 3 would be .46, .39, and .53 respectfully. The floor percent correct scores are different for each judge depending on how severe each judge rated taking no action in each scenario. For example, one judge may have seen taking no action as very endangering of the safety of the flight and rated no action a one, while another judge rated the situation as less severe and rated a three.

Scenario Breakdown by Category

To understand where the difference between the high-time and low-time group lay, analyses were conducted on each of the three scenario categories: Memory, Declarative Knowledge, and Spatial. For example, if low-time pilots performed extremely poor on problem scenarios highly dependent upon spatial abilities, but were no worse on scenarios highly dependent upon Declarative Knowledge and Memory, then this would color the overall optimality result found above. Consequently, mean optimalities for each scenario type were calculated and t-tests performed.

Each scenario had been rated as low, medium, or high in each of three categories in a previous study (Wickens et al., 1988). The scenarios rated as "high" in a Category were used to investigate any group differences. The problem scenarios could be placed into one or more of the categories if they had been rated "high" in that realm. As a result, high ratings were given to nine Memory, fourteen Declarative Knowledge, and five Spatial scenarios. It should be noted that every problem scenario requires the use of all ability categories to some extent, but the evaluation is only examining the most prominent demand. For example, a problem rated "high" in Declarative Knowledge demand, and "medium" in Spatial demand, is only scored under the Declarative Knowledge Category.

The judges' scores were combined and an overall optimality percentage was calculated for each group in each of three categories. Interestingly, high-time pilots performed significantly better than the low-time pilots in all three areas. Table 6 lists the means and significance level for each Category.

Table 6.

Open-Ended Response Optimalities by Scenario Category

	High-time Pilots	Low-time Pilots	Significance
Memory	.6662	.5077	t (24) = 7.01; p < .001
Declarative Knowledge	.7038	.5577	t (24) = 6.25; p < .001
Spatial Awareness	.6077	.4477	t (24) = 4.96; p < .001

Both groups of pilots performed best on problems rated highly dependent upon Declarative Knowledge, second best on Memory scenarios, and performed worst on Spatial scenarios. Nonetheless, the differences in each category are dramatic with the high-time group scoring approximately 15-16% better. The difference in Declarative Knowledge is somewhat surprising given the results of the Declarative knowledge pretest. High-time and low-time pilots were essentially equal on a multiple choice format pretest of Declarative knowledge concerning instrument flight rules and regulations. Applying this knowledge in a real-time situation, however, appears to be much more difficult for the low-time pilots. Perhaps this phenomenon was also due to the prompting element of multiple choice type questions being removed.

The difference in Memory scores is consistent with the hypothesis that low-time pilots are forced to integrate cues in working memory, whereas the high-time pilots were able to match the cues with a familiar situation in LTM. Under the Stokes (1991) model, both groups of pilots attempt to find a pattern match in LTM--the low-time pilots simply fail more often than the high-time pilots. Under this premise, low-time pilots are expected to perform less optimally because they are forced to generate a unique solution using a working memory fallible to time pressure and perceived stress. The high-time pilots, on the other hand, were able to rely on matching the cues with a script from past experiences and quickly recall a "tried and true" solution.

As predicted, the difference in the spatial awareness category was enlightening. The fact that high-time pilots were no different from low-time pilots in the non-domain specific arena, but were significantly better on the domain specific task suggests spatial problem solving follows the Stokes (1991) model as well. Presumably, spatial templates could be recalled from LTM for familiar situational schemata, and high-time pilots have these spatial schemata more readily available.

Situation Cue Recognition

It was hypothesized that the source of inferior decision making in low-time pilots was rooted in poorer cue recognition. Therefore low-time pilots would report fewer correct cues and more irrelevant cues when analyzing a situation than high-time pilots. High-time pilots did, in fact, identify significantly more relevant cues and reported fewer irrelevant cues than low-time pilots. Table 7 lists the means and p-values.

Table 7.

MIDIS Performance -- Relevant and Irrelevant Cues

	Relevant Cues	Irrelevant Cues
High-time pilots	15.38	9.92
	[t (24) = 4.20, p<.001]	[t (24) = 1.21, p<.237]
Low-time pilots	8.92	12.77

Alternative Selection Differences

It was hypothesized that the satisficing strategy of high-time pilots would result in low-time pilots generating more alternatives than high-time pilots, because high-time pilots will use their first workable alternative and low-time pilots will consider the utility of many alternatives. However, the opposite result was found. High-time pilots generated significantly more alternatives than low-time pilots (mean scores 30.38 and 21.69, respectively; t (24) = 2.32, p < .029).

The fact that the MIDIS flight prompted the subjects to list all alternatives a reasonably competent pilot would consider may have confounded this result. High-time pilots were better able to comply with the directions of the experiment and list more alternatives, although in an actual flight they may not have attempted to think of more possibilities once a satisfactory alternative was found. Therefore, simply summing the total number of alternatives listed would not identify the hypothesized decision process.

The satisficing strategy was also posited to lead high-time pilots to choose the first alternative generated as the most optimal solution more often than low-time pilots. Scoring the percentage of first alternatives chosen including when only one alternative was listed resulted in high-time pilots choosing their first alternative 79 percent of the time, while low-time pilots chose their first alternative 68 percent of the time. Dissecting this finding revealed the hypothesized phenomenon. When multiple alternatives were considered, high-time pilots chose their first response 71 percent of the time, while low-time pilots went with their first response only 53 percent of the time. Although this difference is impressive, it actually underestimates the extent to which high-time pilots actually choose their first response relative to the low-time group. The high-time group cited a much larger number of possible solutions (116 vs. 72), than the low-time group. And therefore had a greater chance of picking an alternative other than the first considered.

For example, a low-time pilot who only listed multiple alternative one time and happened to chose the first alternative went 1 for 1 and received a percentage score of 100%. A high-time pilot, on the other hand, listed multiple alternatives in ten situations and chose the first alternative all ten times. This person went 10 for 10 and was scored a 100%. The high-time pilot's raw data shows a much greater persistence towards choosing the first alternative than the low-time pilots. Simply scoring by percentages, therefore, presents a false picture of each group's behavior. Accordingly, a weighting scheme was developed to correct this artifact.

Appendix C gives a detailed explanation of the weighting scheme. Basically, individual scores were weighted by reducing the proportion of first alternatives chosen from the 25 possible scenarios in which multiple alternatives could be listed. This difference between groups was consistent with the hypothesis that high-time pilots would choose the first alternative generated as the most optimal solution more often than low-time pilots [mean scores 3.72 (high-time) and 2.01 (low-time); $t\ (24) = 3.37$; $p < .003$]. Table 8 lists the number of times first alternatives were chosen over the number of times more than one alternative was considered, as well as each score's corresponding weight.

Table 8.

MIDIS Performance -- First Alternative Chosen as Best Listed

High-time pilots		Low-time pilots	
1st Alternative picked # situations multiple alternatives considered	Weight	1st Alternative picked # situations multiple alternatives considered	Weight
5/13	2.40	6/7	4.32
2/4	1.68	5/10	3.00
1/3	.88	2/5	1.60
1/1	.96	5/8	3.40
4/6	3.04	5/6	3.80
3/5	2.40	8/10	4.80
4/6	3.04	7/11	3.92
8/11	4.48	6/10	3.60
3/7	2.16	1/5	.80
4/7	2.88	7/8	4.76
3/6	2.28	6/9	3.84
0/1	.00	10/10	6.00
0/2	.00	14/17	4.48
38/72	Mean = 2.01	82/116	Mean = 3.72

Mean% = .53 | **Mean% = .71**

Correlational Analyses of MIDIS Data

Correlational analyses were performed next to examine the relationships between the significant variables. The ATC Recall task, ATC Situation Recognition task, and Number of Relevant Cues detected were all significantly correlated with MIDIS flight open-ended decision optimality. For illustrative purposes judge 2 decision optimality is plotted against each variable. Judge 2 was chosen because he had the highest combined inter-rater reliability values with the other two judges.

ATC Situation Recognition vs Decision Optimality

Figure 3 presents the relationship between the ATC Situation Recognition task score and open-ended Decision Optimality. High-time pilots are coded with the number 2, while low-time pilots are coded with 1. The correlation (r=.6517) is significant at the p < .001 level. The relationship clearly shows that pilots who scored well on the ATC Situation Recognition task performed more optimally on the MIDIS flight open-ended decisions. The plot also reveals two distinct groups. High-time pilots are concentrated in the upper right corner, while low-time pilots are concentrated in the lower left. This is consistent with the hypothesis that high-time pilots are better able to use LTM schemata, and consequently perform better than pilots less able to recall scripts from LTM.

Number of Relevant Cues Detected vs Decision Optimality

Number of Relevant Cues Detected is plotted against Decision Optimality in Figure 4. It was hypothesized that high-time pilots would detect a greater number of relevant cues, which would lead to higher optimalities in decision quality. This hypothesis is strongly supported by the correlation (r=.8014, p < .001). As shown in figure 4, the relationship is very linear with the two cohort groupings very distinct.

ATC Recall vs Decision Optimality

ATC Recall task concept difference scores were then plotted against Decision Optimality in Figure 5. Since this task was designed to be sensitive to situational schemata, it was hypothesized that high-time pilots would perform better on this task which would in turn lead to higher decision optimality. The correlation between ATC Recall and Decision Optimality is significant (r=.3393, p < .045), but not as strong as the previous findings. Nonetheless, it further supports the LTM strategy hypotheses. It appears that the high-time pilots' ability to easily recall domain-dependent (aviation relevant) information in the operational context, enhances their decision optimality.

ATC Situation Recognition vs Number of Relevant Cues

Correlations between the predictor variables were then calculated and plotted against each other. Figure 6 depicts the relationship between ATC Situation Recognition scores and Number of Relevant Cues detected. This correlation was significant (r=.3851, p < .026) suggesting that pilots able to effectively access LTM schemata for various types of situations are better able to correctly recognize relevant cues pertinent to the problem.

Figure 3.

ATC RECOGNITION TASK AND DECISION OPTIMALITY

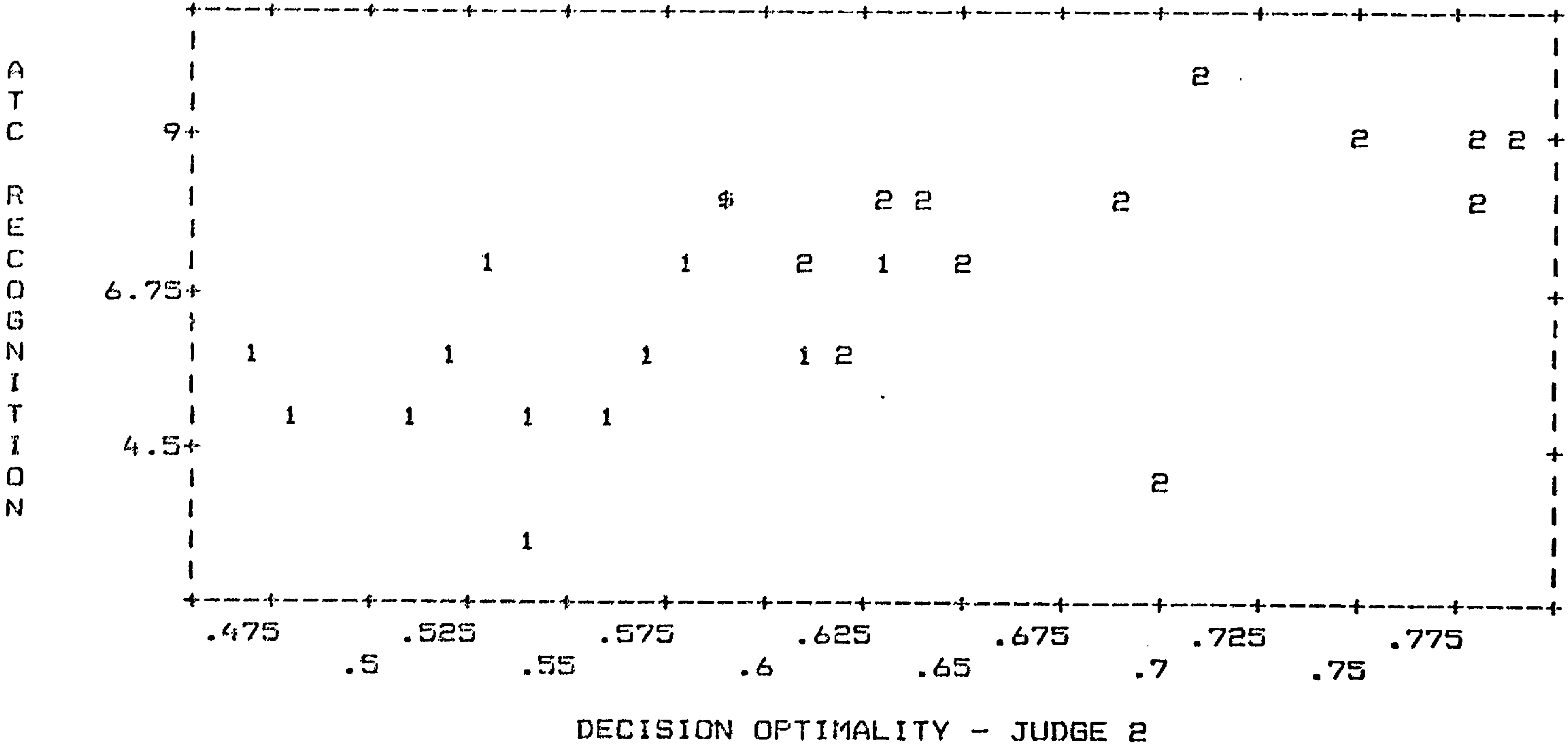


Figure 4.

DECISION OPTIMALITY AND NUMBER OF RELEVANT CUES DETECTED

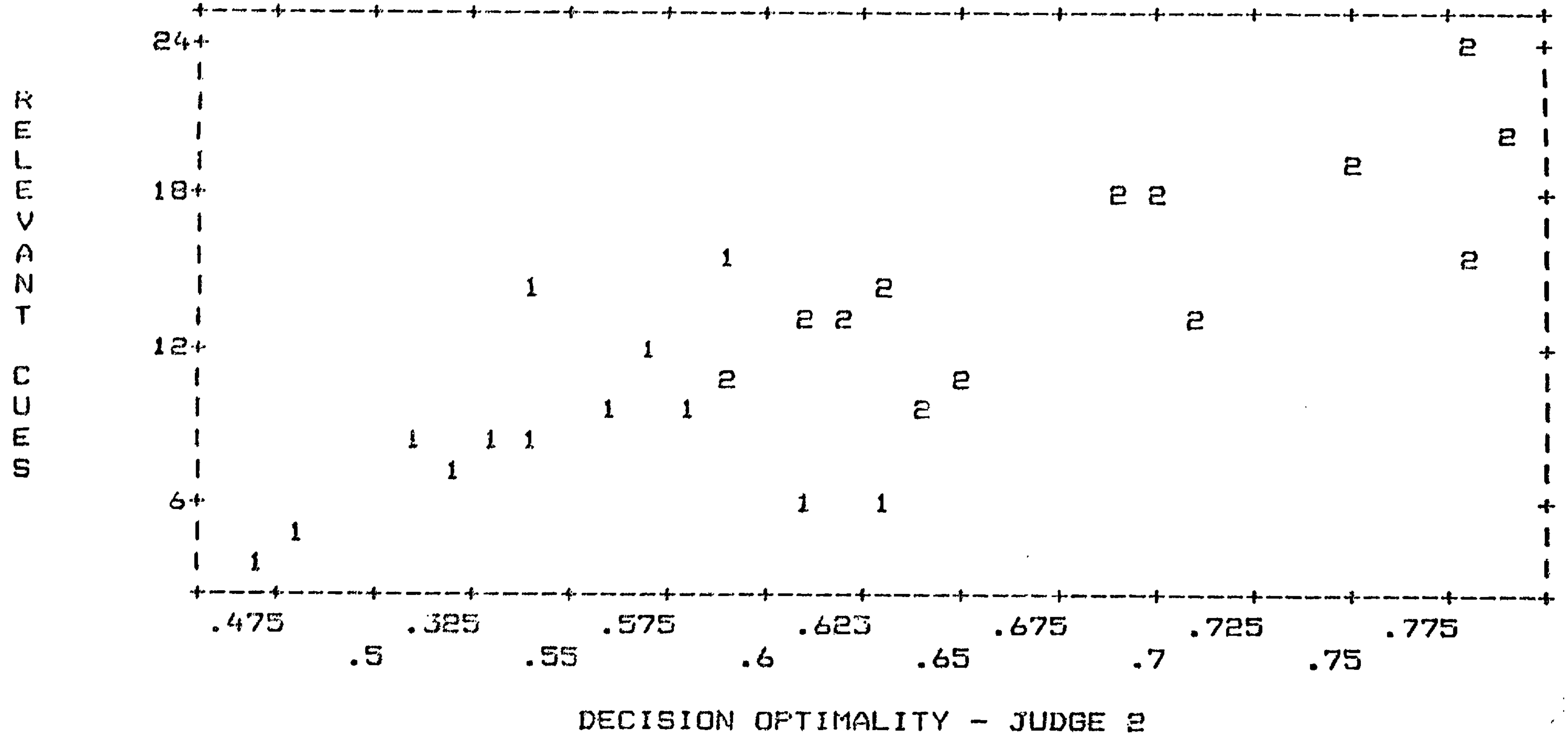


Figure 5.

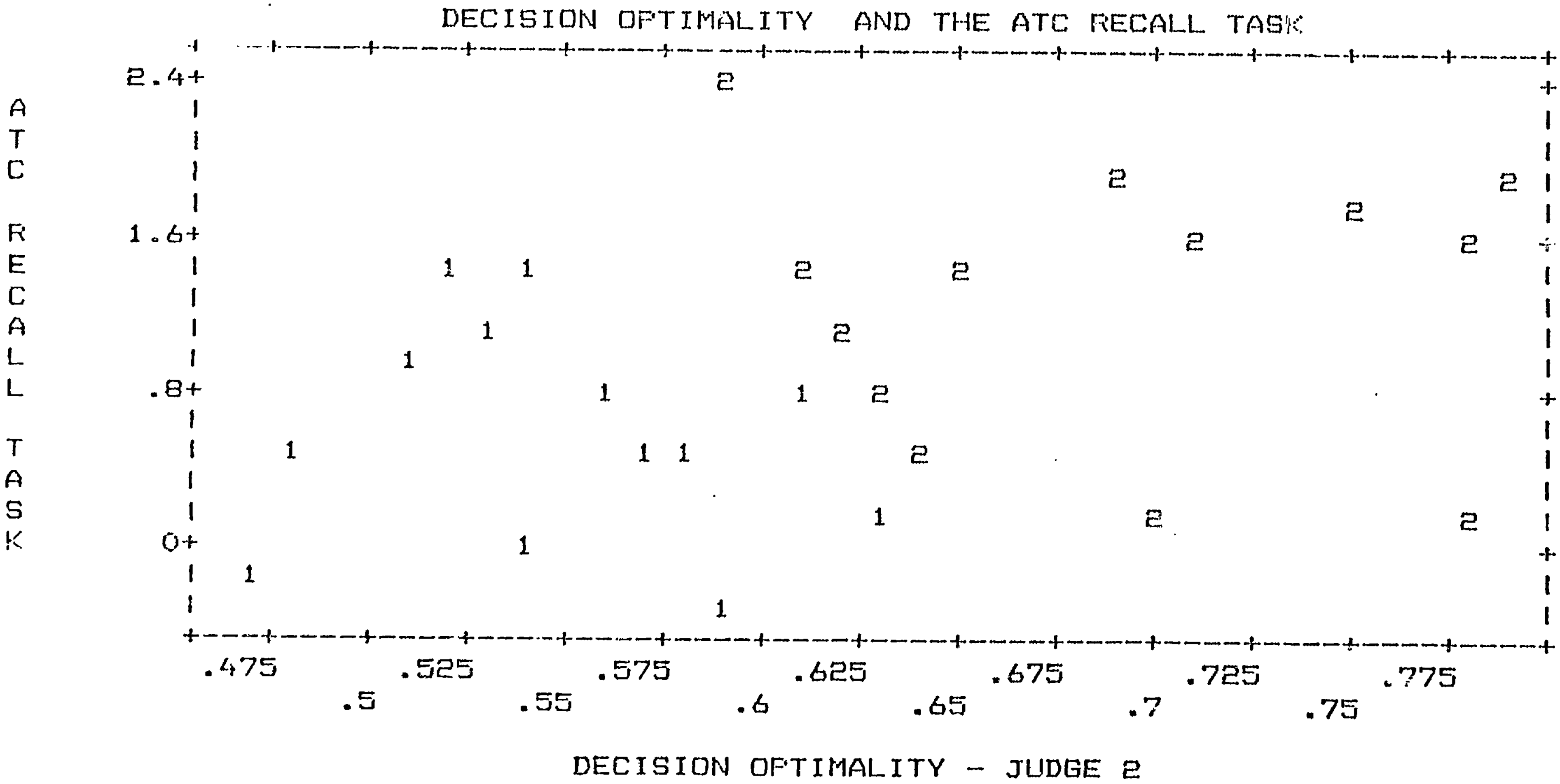
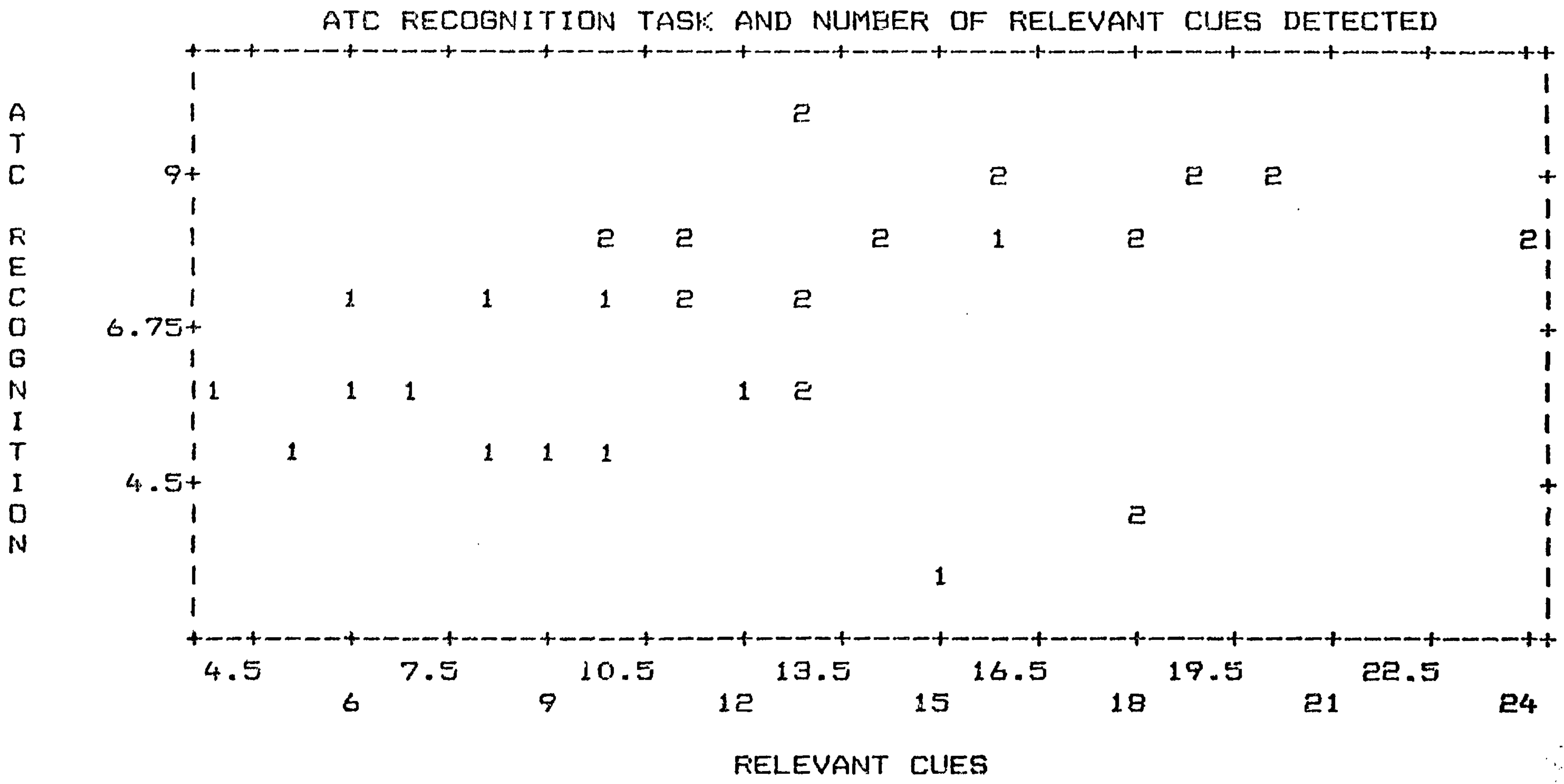

DECISION OPTIMALITY AND THE ATC RECALL TASK
ATC RECALL TASK
2.4
1.6
.8
0
.475
.5
.525
.55
.575
.6
.625
.65
.675
.7
.725
.75
.775
DECISION OPTIMALITY - JUDGE 2

Figure 6.


ATC RECOGNITION TASK AND NUMBER OF RELEVANT CUES DETECTED
ATC RECOGNITION
9
6.75
4.5
4.5
6
7.5
9
10.5
12
13.5
15
16.5
18
19.5
21
22.5
24
RELEVANT CUES

ATC Situation Recognition vs ATC Recall

ATC Situation Recognition score was also significantly correlated with the ATC Recall task's concept difference score. Figure 7 illustrates the significant correlation ($r=.3845$, $p < .026$), verifying the relationship between the two LTM tests. These tasks were expected to be significantly related, because they are both measuring access to LTM strategies. However, it is important to

realize that the shared variance (r^2) is only 14 percent; thus suggesting each test is measuring something unique by itself.

Number of Relevant Cues vs ATC Recall

Interestingly, the relationship between Number of Relevant Cues detected and the ATC Recall task was not significant ($r=.2545$, $p < .105$). This result further indicates that the ATC Recall task is measuring a separate LTM ability than the ATC Situation Recognition task. Recognizing relevant cues is influenced by spatial ability. The ATC Situation Recognition task was designed to measure LTM spatial schemata. The ATC Recall task, on the other hand, does not place as heavy a demand on spatial ability, but rather on verbal scripts for particular situations. The ability to recall situational verbal scripts, (for example, departure instructions), is related to decision performance, but not through correct cue recognition. The fact that the ATC Recall and Number of Relevant Cues variables are not significantly related, implies that the ATC Recall task is measuring a different LTM ability than the ATC Situation Recognition task.

Figure 7.

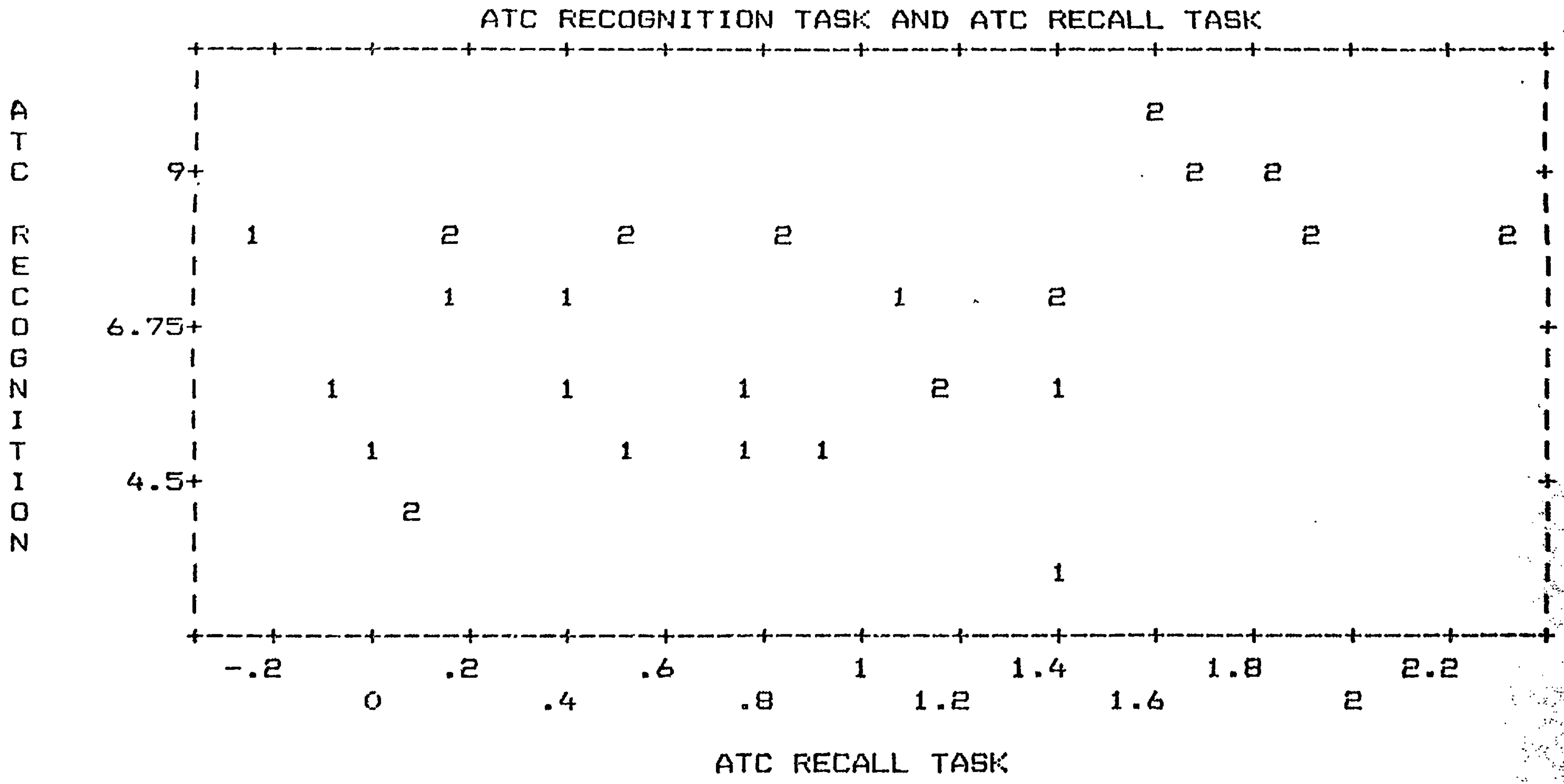
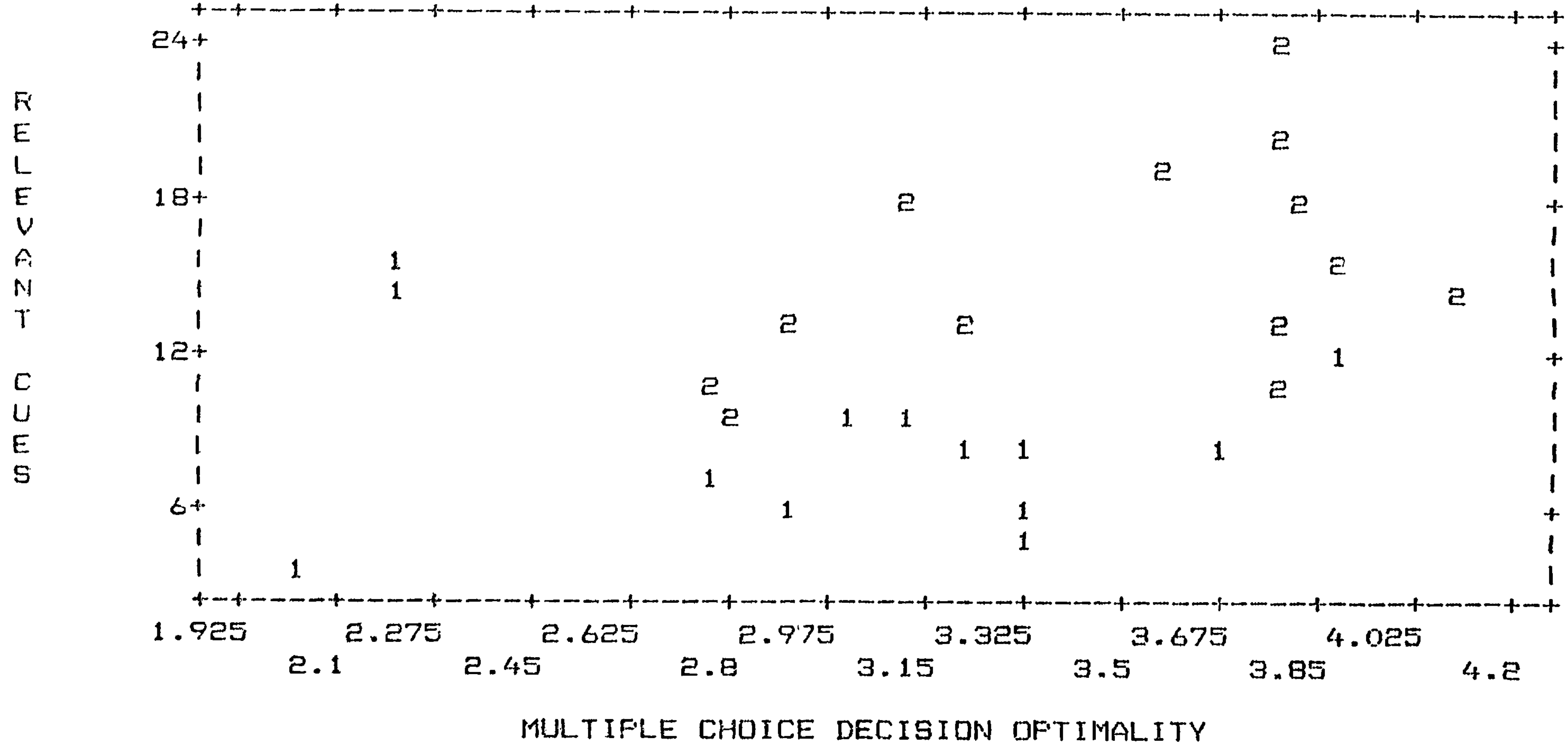

ATC RECOGNITION TASK AND ATC RECALL TASK
ATC RECOGNITION
9
6.75
4.5
-.2
0
.2
.4
.6
.8
1
1.2
1.4
1.6
1.8
2
2.2
ATC RECALL TASK

Figure 8.

DECISION OPTIMALITY AND NUMBER OF RELEVANT CUES DETECTED



Predictor Variables vs Multiple Choice Optimality

Investigating the relationship between the predictor variables and multiple choice decision optimality revealed only one significant relationship. Figure 8 illustrates the significant correlation (r=.3592, p < .036) between the Number of Relevant Cues detected in the open-ended scenarios with multiple choice decision optimality. This finding demonstrates that the high-time groups recognition of the cues relevant to a problem was consistent across situations with the multiple choice question format.

Summary of Correlational Analyses

In sum, the ATC Situation Recognition task, ATC Recall task, and Number of Relevant Cues Detected are all significantly correlated with open-ended decision optimality. Furthermore, the plots highlight the differences between the two groups of pilots. The results of the correlational analysis are consistent with hypothesis that when presented with a situation which could endanger the safety and/or efficiency of a flight, high-time pilots draw upon long term memory based situational schemata and use these scripts as templates to fill in new information from relevant cues. This process leads high-time pilots to implement safer, more efficient solutions. Low-time pilots, on the other hand, are not able to draw upon situational schemata in LTM as readily. This forces them to search for, and integrate cues in working memory using real-time inferential processes. As a result, they are not able to generate alternatives which are as safe and/or efficient as their more experienced colleagues.

Within-Group Analyses

The previous analysis explored the differences existing between the two groups of subjects, the high-time and the low-time pilots. It remains to be determined, however, which, if any, of the psychometric variables predict performance on the MIDIS simulated instrument flight. In particular, we need to ascertain whether information processing capabilities tests are better at predicting operational performance, or whether long term memory tests are more useful. These questions will be addressed using multiple regression analysis.

Stepwise Multiple Regression

The stepwise multiple regression procedure terminates with the identification of a single "best" regression model (Neter, Wasserman, Kutner, 1990), which in this procedure, the first variable considered for entry in to the prediction equation is the one with the largest absolute correlation with the dependent variable. If this variable passes the entry criterion, the F-test, then the second variable is selected based on the highest partial correlation. From this point on, each variable in the equation is examined for removal and variables not in the equation are examined for entry. This process terminates when no more variables meet the specified entry and removal criteria (Norusis, 1986).

The general strategy was to begin with the very broad application of the SPSS/PC+ stepwise multiple regression sub-program, then focus in on the different types of predictor variables (i.e. biographical information, knowledge representation measures, psychological variables, etc...). The judges were kept separate to incorporate a replication. Even though the inter-rater Reliabilities were high enough to justify the combining of the independent, "blind" judges' ratings into a single overall dependent variable, "Optimality," each set of judge's ratings examined separately in what amounts to "built-in" replications. Utilizing this design provided a very stringent "triple jeopardy" experiment, as it were, which permits the evaluation of robustness of the results, or conversely, how judge specific the results were. The strategy adopted initially included all variables regressed against each judge's Decision Optimality, then the biographical information was taken out, and then criterion task based measures were excluded from the analysis. This process continued progressively paring down the variables until only the psychological variables plus Age (which was kept in as a control) remained.

Multiple Regression With All Variables: General Approach

The first analysis regressed all psychological, biographical, and performance variables on the optimality ratings from each judge. The stepwise multiple regression technique identified many significant predictors of decision optimality. In view of this, the first five have been listed in Tables 9-11, but only the first three predictors entered should be considered for inference. The rule of thumb when determining the reliability of predictors is ten subjects per variable (Alster, 1991). The fourth and fifth variables are interesting to note, but due to the sample size relative to the number of predictor variables, they are less reliable. Nonetheless, important information may be gained by examining patterns of predictors in relation to the dependent variables.

Table 9 presents the results of the first regression analysis for each judge. Listed down the left margin are the predictor variables in order of importance as selected by the stepwise procedure, along with the total variance accounted for. This variance (and the associated significance level) has been adjusted, that is, corrected downward using an SPSS/PC algorithm to offset the potential capitalization on chance associated with multiple regression analysis using many independent variables (Tatsuoka, 1976).

Results

The single most powerful predictor of decision performance was the "certificate held by the pilot". What this means is that a pilot holding an airline transport rating, for example, tends to achieve higher optimality scores than a pilot with just an instrument rating. This makes intuitive sense because each rating is supposedly a measure of competence or expertise. Stringent written, oral, and flight examinations must be passed before a higher rating is received. The finding is not surprising, but it is important because the "Level of Certification" variable was able to account for over half of the total variance in decision optimality. Moreover, the result appears to be a robust one: "Level of Certification" accounts for the most variance in all three independent judges. Between 55.6 and 64.9 percent of the variance in pilot decision making can be accounted for simply by knowing the pilot's flight qualifications.

It is also important to note that it was the Level of Certification held by the pilot which was the best predictor of performance, and not Total Hours of flight experience. In fact, the variable Total Hours was never selected in any of the regression analyses performed. This is consistent with the definition of pilot expertise, in that expertise is a reflection cf the **quality** of flight time, which can be measured by the pilot's level of certification, as opposed to the shear **quantitative** measure of time in the sky.

The second best predictor variable for the optimality ratings of judge 1 was the percentage of time the first alternative was chosen by the subject pilot. This result implies that those pilots who most often acted on the first alternative they considered, performed more optimally than those who did not. This is potentially an important finding insofar as it is consistent with the premise that high performing, experienced pilots use a satisficing strategy in which the first workable alternative is chosen from long term Memory based schemata. However, the finding is not "replicated" with the other two judges.

Relevant Cue Percentage was the second predictor variable for judge 2. This result implies that pilots who detect more cues relevant to the problem, select a more optimal alternative to deal with the situation. This finding is in contrast with Simmel et al. (1987), who argue that even though correct cue diagnosis has

Table 9.

Multiple Regression Analysis (N=26) -- All Variables

Judge 1

Predictor	Adjusted R^2	(Sig)*
1. Certification Held by Pilot	.584	(.001)
2. First Alternative Total % Chosen	.702	(.001)
3. 1st order tracking single-dual task time decrement	.766	(.001)
4. Non-target Sternberg % correct responses	.798	(.001)
5. Backwards Visual # Span Incorrect Response Latency	.844	(.001)

Judge 2

Predictor	Adjusted R^2	(Sig)*
1. Certification Held by Pilot	.649	(.001)
2. Relevant Cue %	.742	(.001)
3. Backwards Visual # Span Latency Correct Responses	.822	(.001)
4. ATC Situation Recognition Responses Correct	.915	(.001)
5. Multiple Choice Optimality	.945	(.001)

Judge 3

Predictor	Adjusted R^2	(Sig)*
1. Certification Held by Pilot	.556	(.001)
2. Backwards Visual # Span Latency Correct Responses	.695	(.001)
3. Dual Task Target Sternberg Latency	.761	(.001)
4. Hidden Figures Incorrect Response Latency	.816	(.001)
5. 1st order tracking single-dual task time decrement	.858	(.001)

* All significance values are preceded by (p <)

occurred, accidents may result due to the overassessment or underassessment of the seriousness of a given consequence. Relevant Cue Percentage was not stable across all three judges, however, indicating further research may be needed to clarify this disagreement.

Latency for correct responses in the Backwards Visual Number Span task (a measure of spatial imaging ability) was the second predictor variable for judge 3. It suggests that subjects who take **longer** to type digits they have seen in the backwards visual number span task perform more optimally. This variable is influenced by the age of the subjects, because older pilots tend to have slower response times on this variable (Stokes, Banich, Elledge, and Ke, 1988).

Multiple Regression Without Certification Level

The second multiple regression was performed on the data without the variable "certification held by the pilot" included to see which information was most predictive if the Level of Certification was not known. Table 10 presents the results of this regression analysis for each judge.

The single most predictive variable when Level of Certification was not included in the stepwise multiple regression, was Relevant Cue Percentage. Again, this result implies that pilots who detect more cues relevant to the problem select a more optimal alternative to deal with the situation. A notable aspect of this finding was that Number of Relevant Cues Detected was, unaided, able to account for around half of total variance, making it a powerful predictor of decision performance.

Table 10

Multiple Regression Analysis (N=26) -- Variables except

CERTIFICATION

Judge 1

Predictor	Adjusted R^2 (Sig)*
1. Relevant Cue %	.495 (.001)
2. First Alternative Total % Chosen	.604 (.001)
3. ATC Situation Recognition Responses Correct	.703 (.001)
4. Risk Taking First Trial	.790 (.001)
5. Visual Conditions Flight Time in Past 90 Days	.836 (.001)

Judge 2

Predictor	Adjusted R^2 (Sig)*
1. Relevant Cue %	.627 (.001)
2. ATC Situation Recognition Responses Correct	.761 (.001)
3. Spatial Memory Latency for Previously Seen Figures	.905 (.001)
4. Dual Task Sternberg Latency for Target Probes	.932 (.001)
5. % 1st Alternative Chosen Out of Multiple Alt. listed	.958 (.001)

Judge 3

Predictor	Adjusted R^2 (Sig)*
1. Relevant Cue %	.403 (.001)
2. Stroop Monochrome to Color Time Decrement	.554 (.001)
3. Concept Difference Recalled from ATC Messages	.622 (.001)
4. Hidden Figures Incorrect Response Latency	.691 (.001)
5. Logical Reasoning % Correct Responses	.766 (.001)

* All significance values are preceded by (p <)

"ATC Situation Recognition responses correct" was selected third for judge 1 and second for judge 2. This variable was the number correct on the ATC Situation Recognition Task testing long term memory schemata. "Concept Difference Recalled from ATC Messages" was selected third for judge 3 and was another variable attempting to measure long term memory schemata. It represents the difference score between the number of concepts recalled in coherent messages minus the number of concepts recalled from random messages. Presumably pilots able to recall situational scripts would perform better than those who could not. It was hypothesized that subjects who scored well on this task were better able to tap long term memory scripts from past experiences. These subjects were hypothesized to perform better than those having to rely on integrating cues in working memory and deriving unique solutions. These results lend support to the aforementioned hypotheses.

"First Alternative Total % Chosen" appears again as the second predictor for judge 1. As discussed above from Table 9, this supports the satisficing strategy used by experienced pilots.

Multiple Regression With Only Psychological Variables Plus Age

Table 11 shows the results of the multiple regression using only the psychological variables plus Age as predictors; Certification held, Number of Relevant Cues Detected and all biographical information were removed. Once again, the common predictor variable among all three judges' was the "ATC Situation Recognition responses correct" score.

Table 11

Multiple Regression Analysis (N=26)--Psychological Variables & Age

Judge 1

Predictor	Adjusted R^2	(Sig)*
1. Spatial Memory Total % Correct Responses	.459	(.001)
2. ATC Situation Recognition Responses Correct	.600	(.001)
3. Spatial Memory Latency for Previously Seen Figures	.721	(.001)
4. Rotated Hidden Figures Latency Incorrect Responses	.758	(.001)
5. Non-target Sternberg Dual Task % Correct Responses	.803	(.001)

Judge 2

Predictor	Adjusted R^2	(Sig)*
1. ATC Situation Recognition Responses Correct	.401	(.001)
2. Spatial Memory Latency for Previously Seen Figures	.672	(.001)
3. Spatial Memory % Correct Previously Unseen Figures	.817	(.001)
4. Spatial Rotations % Correct Responses	.836	(.001)
5. Visual Number Span % Correct Responses	.864	(.001)

Judge 3

Predictor	Adjusted R^2	(Sig)*
1. Spatial Memory % Correct Previously Unseen Figures	.323	(.001)
2. Spatial Memory Latency for Previously Seen Figures	.528	(.001)
3. ATC Situation Recognition Responses Correct	.629	(.001)
4. Spatial Rotations Latency for Correct Responses	.717	(.001)
5. No Variables Selected	---	---

* All significance values are preceded by (p <)

The other variables were all predominantly spatial ability measures. For example, "Spatial Memory Total % Correct Responses" and "Spatial Memory % Correct Previously Unseen Figures" were calculated from the spatial memory task. This task tested the subject's ability to recall whether they had previously seen abstract shapes before. Those subjects correctly recognizing shapes they had seen, as well as shapes they had not seen before, performed more optimally on the MIDIS flight. Although the two groups failed to differ significantly on domain-independent spatial abilities, these abilities become important in performance prediction among only the other psychological variables. This result is consistent with the belief that aviation, in general, is laden with spatial demands.

Multiple Regression Within Each Group Separately

The ensuing investigation examines the pattern of predictors within each group analyzed Separately. Table 12 lists the values for low-time pilots; Table 13 lists values for high-time pilots. The first three predictors, the adjusted variance accounted for, and corresponding significance levels are presented for each group. Due to decreased sample size, the second and third variables listed should not be considered as reliable as the first variable.

Within the low-time pilot group, performance was predominantly predicted by information processing spatial ability and focused attention measures. Within the high-time pilot group, information processing spatial ability variables were the dominant predictors. "Risk Consistency" was the second best predictor for judge 3 implying that subjects who consistently took the same amount of

Table 12

Multiple Regression Analysis for Low-Time Pilots (N=13)

Psychological Variables & Age

Judge 1

Predictor	Adjusted R^2 (Sig)*
1. Spatial Memory Total % Correct Responses	.430 (.009)
2. Zero Order Tracking Time Off Target	.660 (.002)
3. Spatial Rotations % Correct Responses	.783 (.001)

Judge 2

Predictor	Adjusted R^2 (Sig)*
1. Non-Target Sternberg % Correct Responses	.279 (.037)
2. Logical Reasoning Latency for Correct Responses	.683 (.001)
3. Visual Number Span Latency for Incorrect Responses	.860 (.001)

Judge 3

Predictor	Adjusted R^2 (Sig)*
1. Backwards Visual # Span Latency Correct Responses	.417 (.010)
2. Target Sternberg Dual Task Latency	.784 (.001)
3. Spatial Memory Incorrect Latency Prev. Seen Figures	.875 (.001)

* All significance values are preceded by (p <)

Table 13

Multiple Regression Analysis for High-Time Pilots (N=13)

Psychological Variables & Age

Judge 1

Predictor	Adjusted R^2 (Sig)*
1. Spatial Memory Latency for Total Correct Responses	.347 (.020)
2. Spatial Rotations % Correct Responses	.507 (.012)
3. Spatial Memory % Correct Previously Unseen Figures	.642 (.006)

Judge 2

Predictor	Adjusted R^2 (Sig)*
1. Spatial Memory Total Latency for Prev. Seen Figures	.331 (.023)
2. Spatial Rotations % Correct Responses	.460 (.018)
3. Spatial Memory Total % Correct Responses	.734 (.002)

Judge 3

Predictor	Adjusted R^2 (Sig)*
1. Spatial Rotations Latency for Correct Responses	.280 (.037)
2. Risk Consistency	.453 (.019)
3. Logical Reasoning % Correct Responses	.655 (.005)

* All significance values are preceded by (p <)

risk performed better than those with greater variance in risk taking.

An interesting finding from this analysis was the amount of variance the domain-independent psychological variables were able to account for in each group. It was hypothesized that low-time pilots would be forced to rely on working memory based abilities more often than high-time pilots. *Ex hypothesi*, the STM information processing variables, which measure working memory capabilities, could be expected to be somewhat predictive of low-time pilots' performance, but less predictive for high-time pilots. When the psychological variables were regressed against performance for each group Separately, the STM information processing variables did, in fact, account for more variance in low-time pilots, than in high-time pilots. Specifically, they accounted for an average 37.5 percent of the variance (averaged over three judges) for the first variable selected. This contrasts with only 31.9 percent of the variance accounted for in the high-time group. When the first three predictor variables are considered, this result is strengthened. Information processing variables accounted for an average 83.9 percent of the variance in low-time pilots, versus only 67.7 percent in high-time pilots, that is, about 25% less.

Summary of Multiple Regression Analyses

In summary, the pattern of data reveals robust and noteworthy trends. The best predictor of pilot performance in simulated instrument conditions is the FAA pilot certificate held. Pilots with advanced certificates (ATP and Flight Instructor) perform better than pilots with less advanced certificates (private license and instrument certificate). This measure alone was so strong, that it was able to account for well over half of the variance in decision optimality. This result contrasts with that found for flight hours, which, although the most ubiquitous index of pilot expertise, appeared nowhere in any regression analysis performed.

If knowledge of the certification held by the pilot was excluded, the best predictor of performance was the Number of Relevant Cues Detected in problem scenarios during the flight. This variable alone was able to account for about half of the total variance (depending on the judge). If biographical and flight performance based measures were excluded, LTM knowledge-based representation measures were the most predictive variables. Generally, the ATC Situation Recognition task, ATC Recall task, and Percentage of First Alternatives chosen measures were selected most often. Next, the regression was confined to psychological variables only, plus Chronological Age. At this point ATC Situation Recognition, the domain-dependent ability measure of LTM dominated the prediction of decision optimality. Finally the high and low-time groups were analyzed separately finding domain-independent spatial ability measures to be the best predictors of decision optimality.

It is important to note that the variable Age was never selected as one of the top five predictors, even though it was retained as a variable in every regression analysis. This suggests that the observed performance differences were not merely a "seasoning" effect, to use the FAA's term. There is no evidence that better decision making inherently comes with advancing years.

Summary of All Results

The study first compared high-time pilots against low-time pilots on domain-independent information processing abilities and found no significant differences. The two cohorts were then compared on domain-dependent knowledge-based representations in LTM; high-time pilots were significantly better than low-time pilots on these measures. Open-ended response decision optimalities were then judged by three experts and compared across judges: inter-rater Reliabilities were very high. The criterion variable (decision optimality), was kept separate for each judge, however, to provide a "built-in" replication. Analyses on performance results from the MIDIS flight found high-time pilots performed significantly better than low-time pilots on open-ended response problem scenarios. Additionally, the high-time group significantly outperformed the low-time group on scenarios composed of multiple choice alternative selection.

Further analyses were directed toward determining why high-time pilots outperformed their counterparts. Analyses found that high-time pilots identified significantly more relevant cues and chose their first alternative considered as the most optimal solution more often than low-time pilots. Correlational analyses were then performed to discover the relationships between the prominent variables. Significant correlations were found between ATC Situation Recognition task and Decision Optimality, Number of Relevant Cues Detected and Decision Optimality, ATC Recall task and Decision Optimality, ATC Situation Recognition and Number of Relevant Cues Detected, and ATC Situation Recognition and ATC

Recall task. The Number of Relevant Cues Detected measure was not significantly correlated with the ATC Recall task.

The final analysis sought to find the best predictor variables for decision optimality. The stepwise multiple regression technique selected the Level of Certification held by the pilot as the best predictor for all three judges. Certificate was then removed and Number of Relevant Cues Detected became the best predictor for all three judges. Each of these variables accounted for over half of the total variance in decision making. The last multiple regression conducted on both groups together only included the psychological variables plus Age as predictors; that is, Level of Certification, Number of Relevant Cues Detected, and all biographical information were removed. The dominant predictor variable among the judges was ATC Situation Recognition. Finally, multiple regressions were conducted on the groups Separately to determine which variables were predictive within each group. Within the low-time pilot group, performance was predominantly predicted by the information processing spatial ability and focused attention measures. Within the high-time pilot group, performance was predominantly predicted by information processing spatial ability variables.

DISCUSSION

The goal of this study was to elucidate whether if high-time pilots differed from low-time pilots in decision making on a simulated flight in instrument conditions using an open-ended response format. Previous flight decision simulation studies using multiple choice questions had failed to find any significant differences in performance between high-time and low-time pilots (Wickens, et al., 1987; Barnett, 1989; Stokes, et al., 1990). This, of course, is a highly counter-intuitive result. It was hypothesized that the multiple choice format of these studies may have attenuated variance, or may have prompted low-time pilots to search for cues they otherwise may have overlooked in the decision making process. This reasoning was supported by the decision making literature based on post-hoc theorization (Simmel & Shelton, 1987), and intuition-based observations (e.g., Calderwood, Klein, & Crandall, 1988; Crandall & Calderwood, 1989; Crandall & Klein, 1987; Klein et al., 1986; Klein & Thordsen, 1989; and Thordsen, Galushka, Klein, Young, & Brezovic, 1987) this literature did suggest that a distinction existed between "experts" and "novices." The present study **clearly demonstrated the difference experimentally.**

The experiment reported here attempted to discover how and why high-time pilots differed from low-time pilots with respect to cue recognition, hypothesis generation, option selection, and optimality of choice. It was first necessary to contrast high-time and low-time pilots' performance in terms of domain-independent information processing capabilities and domain-specific LTM knowledge representations. These tests uncovered the basic abilities each group brought to the MIDIS 3.0 decision simulation. After the performance differences were found, the focus of the study was to determine which individual difference measures best predicted decision performance. This analysis was performed with the two groups combined, as well as independently.

Based on previous studies of pilot judgment and decision making using the MIDIS simulation (Wickens, et al, 1987; Barnett, 1989; and Stokes, et al, 1990), and guided by a model of decision making presented by Stokes (1991), several hypotheses were generated. It was hypothesized that high-time pilots would not differ from low-time pilots in terms of STM processing capabilities, yet would possess more elaborate aviation-specific LTM knowledge representations than the low-time group. These differences were posited to account for high-time pilots outperforming low-time pilots on the MIDIS flight.

Under the Stokes (1991) model, pilots retrieve domain-specific problem schemata or "scripts" directly from LTM. Only if unable to pattern match the cues in this way do, pilots use an alternative strategy using real-time computational/inferential processes heavily dependent upon working memory. The LTM related strategy was hypothesized to be more readily available to high-time pilots (with their appreciable experiential repertoires) than to low-time pilots, and thus account for performance differences in decision optimality. The outcomes of the experiment were overwhelmingly consistent with these hypotheses. The discussion to follow will highlight the results not expected, and suggest methodological improvements for future research.

Test of LTM Knowledge Based Representations

The ATC Situation Recognition task proved to be an excellent measure of LTM spatial mental models. Additionally, as hypothesized, it served as one of the best predictors of decision optimality. The ATC Recall task was successful, but less so and could be improved before being used in following studies. Even in its present form, however, the ATC Recall measure was able to significantly discriminate the two groups' ability to recall concept differences in coherent versus random radio message sequences. Further, it appears to be measuring a related, yet different, part of LTM based knowledge than the ATC Situation Recognition task. The ATC Recall task also had shortcomings, in that it failed to find differences in the number of words recalled under each condition, and was not predictive of MIDIS performance.

The failure to find a difference in the number of words recalled may be due to the grading technique. Each word recalled from the exact message was given one point. This resulted in function words (i.e. of, the, to, with) being given equal importance as key words. In subsequent research, a predetermined list of function words which should not be counted ought to be devised. Alternatively, a weighting scheme which gives key words more weight than function words may suffice.

Finally, an item analysis should be conducted to determine which radio messages in the ATC Recall task should be revised. It was apparent that some messages were very good at discriminating the two groups, while others were not. The messages poor in separating the groups were those ostensibly coherent messages which happened to be rather complicated and confusing. One example of a confusing statement is "turn left at the third right on taxiway Papa, then left at November." High-time pilots often commented that if given a confusing message like this, they would have asked for clarification. Messages such as these should be revised using actual radio calls which are less confusing.

Alternative Generation

Based on the Stokes (1991) model of pilot decision making, it was hypothesized that low-time pilots would consider, and therefore list, more alternatives than high-time pilots. High-time pilots, presumably, would successfully make pattern matches in LTM from past experiences. The use of these scripts would enable them to generate a workable alternative without having to consider many options. The low-time pilots, on the other hand, would not be able to make pattern matches as often, and would be forced to use working memory to consider the utility of many alternatives. However, the opposite result was found. High-time pilots generated significantly more alternatives than low-time pilots (mean scores 30.38 and 21.69, respectively; t (24) = 2.32, $p < .029$).

However, this finding was probably a methodological artifact which had been anticipated as a possibility. It seems likely that the very structure of the MIDIS simulation and the nature of the testing situation led high-time pilots to better comply with the directions of the experiment and list more alternatives, although in an actual flight they may not have attempted to think of more possibilities once an early satisfactory course of action was determined. Therefore, simply summing the total number of alternatives listed may have been inadequate to identify the hypothesized decision process.

To compensate for this methodological difficulty, a percentage score was devised indicating how often the first response was chosen from the list of multiple responses. High-time pilots chose their first response listed 71 percent of the time, while low-time pilots chose their first response only 53 percent of the time. A weighting scheme was constructed to analyze the data; the difference was very significant. In light of this data, it is convincing that high-time pilots would, in a real operational setting, have "gone with" their first solution more often, and therefore, would have gone on to consider fewer alternatives. Thus, even though the initial hypothesis was not supported by the results in the strict sense, the logic leading up to the hypothesis appears to be correct.

Confidence Measures

The ATC Situation Recognition task, and all MIDIS scenarios were accompanied by a 5-point Likert scale confidence rating. Unfortunately, the scope of this paper coupled with time constraints did not allow this data to be analyzed. However, the confidence data may also contain very enlightening results and should be analyzed. Of particular interest are how confidence and optimality relate to one another, both between and within-groups. Next, the distinction between multiple choice confidence and open-ended response confidence should be investigated. Finally, the relationship between the ATC Situation Recognition confidence measure and performance on that LTM task, and how they relate to confidence and performance on the MIDIS task are of particular interest.

Multiple Choice Optimality

The significant difference found between high-time and low-time pilots on the multiple choice decision scenarios of the MIDIS flight was somewhat surprising. Although the high-time group was expected to choose more optimal solutions, all previous MIDIS experiments had failed to find criterion performance significantly different (Wickens, et al, 1987; 1988; Barnett, 1989; and Stokes, et al, 1990).

Unlike the previous versions of MIDIS where all problem scenarios were multiple choice, the 3.0 version converted all but nine of the scenarios into open-ended response format. The difference found in the current study may be a function of those particular questions left in multiple choice format. These nine questions alone might have always been able to separate the two groups, but when combined with all the other poorer questions, no overall difference was found.

A competing hypothesis is that the two groups were polarized enough in their experience levels, that the multiple choice type questions, although less sensitive than open-ended questions, were also able to discriminate between the groups. This hypothesis is probably the more plausible. The groups in this experiment were categorized more stringently than in the past. This resulted in a larger gap in flight experience (in terms of certification, instrument time, total hours, etc...) than in previous experiments.

Expertise Studies

The results of this experiment are consistent with those of other anecdotal and observational studies of expertise. The evidence of this study supports the belief that pilots do in fact attempt to rely on aviation-specific structural representations in LTM to recall situational schemata when problem solving. Only when unable to make a pattern match using this satisficing strategy, do pilots "drop down to" a utility-based strategy in working memory to generate and evaluate alternatives.

For a detailed example, taken from Stokes et al. (1990), consider the following instrument panel indications: the turn coordinator continues to show level flight, the AI (attitude indicator) begins to show a roll to the right, and the DG (directional gyro) begins to show a turn to the left. The experienced pilot will, without further analysis, immediately recognize a probable suction-pump failure. The novice, on the other hand, cannot go directly from the visible symptoms to the diagnosis, since there is nothing in LTM to "pattern-match" the symptoms to. The novice can only diagnose the problem via a real-time integrative and inferential process which does require some basic declarative knowledge in LTM. Even executed efficiently, the process would have to proceed something like this: the two attitude instruments disagree. The gyro in the AI is vacuum driven. The gyro in the turn coordinator is electric. One of those systems must not be working. The DG is not consistent with either the AI or the turn-coordinator. The gyro in the DG is not consistent with either the AI or the turn-coordinator. The gyro in the DG is vacuum driven. If the electric gyro had failed the AI and DG would agree. The vacuum instruments have both failed at the same time.

They have one component in common - the engine driven suction pump. Therefore, the suction pump must have failed.

This decision process is similar to that of the chess masters, in Chase and Simon's research (1973), who assess the configuration of the chessboard and look for the "high value move". Similar to the findings of this experimental study, Chase and Simon (1973) observed that the differences in chess playing ability was not a function of cognitive abilities, such as working memory capacity or depth of processing. Rather, the distinction between the chess master and the novice was rooted in the differences between their structural knowledge representations.

The results of this study appear to be transferrable to other domains as well. Klein et al.'s observational studies with fireground command, military commanders battle planning, critical care nursing, corporate information management, and speed chess tournament play are all consistent with this study's experimental results. Klein et al. observed that in each of these domains, expert decision makers are less concerned with searching an exhaustive list to find the best alternative, and most often choose the first workable alternative they can think of (Calderwood, Klein, & Crandall, 1988; Crandall & Calderwood, 1989; Crandall & Klein, 1987; Klein et al., 1986; Klein & Thordsen, 1989; and Thordsen, Galushka, Klein, Young, & Brezovic, 1987). The common link between the experimental and observational research is that both involve time critical situations. It appears that these results are generalizable to other domains typically involving time critical decision making.

Implications

The results of this study may have implications in the areas of personnel selection and training. Evidence was offered that expertise, which has traditionally been measured in total hours, that is **quantity** of flight time, is better measured in terms of the quality of flight time. Pilot selection decisions should give greater weight to the Level of Certification, and perhaps less to total flight hours, as the most important indicator of expertise.

The data from this analysis suggests pilot training could be improved by explicitly building the experiential repertoires of students. Ground school training classes should therefore be modified to include or emphasize event-based learning, in conjunction with traditional fact and rule based training. Another possibility is to increase students' LTM schemata through vicarious experience. Flight simulators or computer aided instruction can simulate in-flight emergencies which build the pilot's experiential repertoire in a more efficient, safer method real-life experience.

Future Research

A fascinating possibility is to expand MIDIS into a flight training device. Students could fly each scenario, making decisions in the open-ended format, followed by a lesson on the computer revealing the most optimal solutions and their reasoning. Scenarios could be replayed so students could detect the relevant cues missed the first time. Additional information, such as the regulation number or ground school lesson, could be given for the student to look up scenarios which were poorly handled.

This experiment found very strong, robust results, but it has to be recalled that these were based on 26 subjects. Nevertheless, much was learned about how high-time pilots differ from low-time pilots. The results of this study merit the expense of larger samples for future investigations. These investigations should focus on answering the "whys" of the results. For instance, little is know about how the expert detects relevant cues so well. It has not been identified when LTM based strategies cause incorrect solutions to be recalled. Finally, it is not known what causes variance in high-time pilots judgment.

REFERENCES

Alster, J. (1991). Personal consultation. November 22, 1991 at the University of Illinois' Computer Services Organization.

Banich, M.T, Stokes, A.F., & Elledge, V. (1987). Cognitive function evaluation in medical certification of airmen--a literature review. (Technical Report ARL-88-4/FAA-88-1). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Banich, M.T, Stokes, A.F., & Karol, D. (1990). Cognitive function evaluation in medical certification of airmen--phase b. (Technical Report ARL-90-2/FAA-90-1). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Barnett, B. (1989). Modeling information processing components and structural knowledge representations in pilot judgment. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Buch, G., & de Bagheera, I.J. (1985). Judgment training effectiveness and permanency. In R.S. Jensen & J. Adrion (Eds.), Proceedings of the Third Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Buch, G., & Diehl, A. (1984). An investigation of the effectiveness of pilot judgment training. Human Factors, 26 (5), 557-564.

Calderwood, R., Klein, G.A., & Crandall, B.W. (1988) Time pressure, skill, and move quality in chess. American Journal of Psychology, 101, 481-493.

Caravella, D.A. (1987). The evaluation of pilot judgment during certification flight tests. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Carroll, J.S. (1980). Analyzing decision behavior: The magician's audience. In T.S. Wallsten (Ed.), Cognitive processes in choice and decision behavior. Hillsdale, NJ: Erlbaum.

Chase, W. & Ericsson, K. (1981). Skilled memory. In J. Anderson (Ed.), Cognitive Skills and their Acquisition. Hillsdale, NJ: Erlbaum.

Chase, W. & Simon, H. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Cohen, M.S. (1987). Mental models, uncertainty, and in-flight threat responses by Air Force pilots. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Connolly, T.J., Blackwell, B.B., & Lester, L.F. (1987). A simulator-based approach to training in aeronautical decision making. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Crandall, B., & Calderwood, R. (1989). Clinical assessment skills of experienced neonatal intensive care nurses. Yellow Springs, OH: Klein Associates Inc.

de Bagheera-Buch, G. (1983). Pilot judgment training validation experiment. In R.S. Jensen (Ed.), Proceedings of the Second Symposium on Aviation Psychology. Columbus, OH: The Ohio

State University.

Eksttrom, R.B., French, J.W., & Harmen, H.H. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Education Testing Service.

Edwards, W. (1986). Decision making. In G. Salvendy (Ed.), Handbook of Human Factors New York: John Wiley.

Gentner, D. & Stevens, A.L. (Eds.). (1983). Mental Models. Hillsdale, NJ: Erlbaum.

Gettys, C.F. (1983). Research and theory on predecision processes Technical Report TR 11-30-83). Norman, OK: University of Oklahoma, Department of Psychology.

Giffen, W.C. & Rockwell, T.H. (1987). A methodology for research on VFR flight into IMC. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Giffen, W.C., Rockwell, T.H., & Smith, P.E. (1985). A review of critical in-flight events research methodology. In R.S. Jensen & J. Adrion (Eds.), Proceedings of the Third Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Goldsmith, T., & Schvaneveldt, R. (1985). ACES: Air Combat Expert Simulation. Technical Report, CRL. MCCS-85-34.

Hunt, R.M. (1979). A study of transfer of problem solving skills from context-free to context-specifi^ fault diagnosis tasks. Unpublished master's thesis, University of Illinois, Urbana, Illinois.

Janis, I.L., & Mann, L. (1977). Decision making: A psychological analysis of conflict, choice, and commitment. NY: The Free Press.

Jensen, R.S. (Ed.). (1981). Proceedings of the First Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Jensen, R.S. (1982). Pilot judgment: Training and evaluation. Human Factors, 24, 61-73.

Jensen, R.S. (Ed.). (1987). Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Jensen, R.S., & Benel, R.A. (1977). Judgment evaluation and instruction in civil pilot training (Final Report FAA-RD-78-24). Springfield, VA: National Technical Information Service.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 251-273.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Klein, G.A. (1989). Recognition-primed decisions. In W. Rouse (Ed.), Advances in Man-Machine Systems Research, 5, 47-92. Greenwich, CT: JAI Press, Inc

Klein, G.A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. Proceedings of the Human Factors Society 30th Annual Meeting, 1, 576-580.

Klein, G.A., & Peio, K.J. (1989). The use of a prediction paradigm to evaluate proficient decision making. American

Journal of Psychology, 102(3), 321-331.

Klein, G.A., & Thordsen, M.L. (1989). Recognitional decision making in C organizations. 1989 Symposium on Command and Control Research, pp. 233-249.

Lester, L.F. & Bombaci, D.H. (1984). The relationship between personality and irrational judgment in civil pilots. Human Factors, 26, 565-572.

Lester, L.F., & Connolly, T.J. (1987). The measurement of hazardous thought patterns and their relationship to pilot personality. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Lester, L.F., Diehl, A.E., & Buch, G. (1985). Private pilot judgment training in flight school settings: A demonstration project. Proceedings of the Third Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Livack, G.S. (1983). Pilot judgment training: Past, present and future. In R.S. Jensen (Ed.), Proceedings of the Second Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Lubner, M.E., & Lester, L.F. (1987). A program to identify and treat 'pilot error', particularly, poor pilot judgment. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), The Psychology of Computer Vision. New York:

McGraw-Hill.

Neter, J., Wasserman, W., Kutner, M.H. (1990). Applied Linear Statistical Models. Homewood, IL: Irwin.

Norusis, M.J. (1986). SPSS/PC+ advanced statistics. Chicago, IL: SPSS Inc.

Rasmussen, J. (1981). Models of mental strategies in process plant diagnosis. In J. Rasmussen & W. B. Rouse (Eds.), Human detection and diagnosis of system failures. New York: Plenum Press.

Rockwell, T.H., Giffen, W.C., & Romer, D.J. (1983). Combining destination diversion decisions and critical in-flight event diagnosis in computer aided testing of pilots. In R.S. Jensen (Ed.), Proceedings of the Second Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Schank, R.C., & Abelson, R.P. (1977). Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. HIllsdale, NJ: Erlbaum.

Simmel, E.C., Cerkovnik, M., & McCarthy, J.E. (1987). Sources of stress affecting pilot judgment. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Simmel, E.C., & Shelton, R. (1987). The assessment of nonroutine situations by pilots: A two-part process. Aviation Space, and Environmental Medicine, 58, 1119-1121.

Slovic, P., Fischoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. Annual Review of Psychology, 28, 1-39.

Sternberg, S. (1966). High-speed scanning in human memory.

Science, 153, 652-654.

Stokes, A.F. (1991). MIDIS-A microcomputer flight decision simulator. In E. Farmer (Ed.), Human Resource Management in Aviation (pp. 107-121). Aldershot, UK: Avebury Technical.

Stokes, A.F. (1991). Flight management training and research using a microcomputer flight decision simulator. American Society of Mechanical Engineers.

Stokes, A.F., Belger, A., Zhang, K. (1990). Investigation of factors comprising a model of pilot decision making: part II. anxiety and cognitive strategies in expert and novice aviators. (Technical Report ARL-90-8/SCEEE-90-2). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Stokes, A.F., Bancih, M.T., Elledge, V., & Ke, Y. (1988). Cognitive Function Evaluation in the Medical Certification of Airmen (Technical Report ARL-88-4/FAA-88-2). Savoy, IL: University of Illinois, Aviation Research Laboratory.

Stokes, A.F., Barnett, B., and Wickens, C.D. (1987). Modelling stress and bias in pilot decision-making. Proceedings of the Human Factors Association of Canada, XXth Annual Conference, Montreal, Oct. 14th-17th, 45-48.

Stokes, A.F., & Raby, M. (1989). Stress and cognitive performance in trainee pilots. Proceedings of the Human Factors Society 33rd Annual Meeting, October 1989.

Stokes, A., Wickens, C., Davis, T. (1986). MIDIS - A Microcomputer based Flight Decision Training System. Proceedings of the 28th International Conference on the Association for the Development of Computer-Based Instructional Systems (ADCIS).

Crystal City, Arlington, VA. (Abstract in Proceedings, p. 380).

Stokes, A., Wickens, C., Davis, T., Jr., Barnett, B., Rosenblum, R., & Hyman, F. (1987). A study of pilot decision making using MIDIS - A microcomputer-based flight decision training system. Prcceedings of the Fourth International Symposium on Aviation Psychology. Columbus, OH.

Stone, R.B., Babcock, C.L., & Edmunds, M.S. (1985). Pilot judgment: An operation viewpoint. Aviation, Space, and Environmental Medicine, 149-152.

Tatsuoka, M. M. (1976). Selected Topics in Advanced Statistics: Validation Studies. Champaign, IL: Institute for Personality and Ability Testing.

Telfer, R. (1987). Pilot judgment training: The Australian study. In R.S. Jensen (Ed.), Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Thordsen, M.L., Galushka, J., Klein, G.A., Young, S., & Brezovic, C.P. (1987). A knowledge elicitation study of military planning (KATR-863C-87-08F). Yellow Springs, OH: Klein Associates Inc.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.

Wickens, C.D. (1984). Engineering psychology and human performance. Columbus, OH: Charles Merrill.

Wickens, C.D., Braune, R., Stokes, A.F., & Strayer, D. (1985). Individual differences and age-related changes: Refinement and elaboration of an information processing performance battery with aviation-relevant task structures (Technical Report EPL-85-1). Champaign, IL: University of Illinois, Engineering Psychology Laboratory.

Wickens, C.D., & Flach, J. (1987). Human information processing. In E. Weiner & D. Nagel (Eds.), Human Factors in Modern Aviation. New York: Wiley.

Wickens, C.D., Stokes, A., Barnett, B., & Davis, T., Jr. (1987). A componential analysis of pilot decision making. University of Illinois Aviation Research Laboratory (Final Technical Report ARL-87-4/SCEEE-87-1). Savoy, IL: Institute of Aviation.

Wickens, C.D., Stokes, A., Barnett, B., & Hyman, F. (1988). The Effects of Stress on Pilot Judgment in a MIDIS Simulator. University of Illinois Aviation Research Laboratory (Final Technical Report). Savoy, IL: Institute of Aviation.

APPENDICES

APPENDIX A: Sample ATC Recall Problems

PROBLEM #1: Ordered Call

Diagram: O'Hare Airport

Aircraft Position: 5 Miles South of O'Hare Receiving Landing Instructions

Message: Report overhead at 2000

for a Right turn to Runway 22 Right,

behind a United 727

on a 3 mile final.

Hold short of 27 Right

and say your type.

PROBLEM #2: Randomized Call

Diagram: Chicago TCA Chart Excerpt

Aircraft Position: 15 Miles East of Palwaukee Receiving Landing Instructions

Message: Braking conditions poor.

Extend your downwind.

Additional traffic is a single-engine Cessna

just reported inbound.

Turn right heading 160

as soon as practicable after takeoff.

APPENDIX B
Subject Biographical Information

Pilot Judgment and Decision Making in Simulated IFR Flight

Name ______________________________ Subject Number ______________

Certificates Held: (circle)
a. Private
b. Instrument
c. Commercial
d. ATP
e. Flight Instructor

Total hours __________ Instrument hours _________

Last 90 Days: VFR hours ________ IFR hours ________
(including simulated)

Total hours in ACTUAL Instrument Conditions: _________

Approximate hours in: Light single engine aircraft _______

Light multi-engine aircraft _______

Heavy multi-engine aircraft _______
(over 12,500 lbs)

THANK YOU

APPENDIX C: Sample Weighting Calculations

There were 25 possible scenario in which multiple alternatives could be listed. A weighting scheme was necessary so that a subject who only listed more than one alternative once and chose it (1/1), did not receive the same score as someone who listed multiple alternatives 10 times and chose the first alternative every time (10/10). A subject who chooses their first response 4/6 times equals 67 percent. Another subject chose the first alternative 6/9 times which also equals 67 percent. The person who was 6/9, however, shows a greater persistence towards choosing their first alternative and should carry a higher weighting.

The formula derived to compensate for this phenomena reduces the number of times the subject chooses their first response by the proportion of the number of times out of 25 possible they listed multiple alternatives. This standardizes the score on a 0 - 25 scale. The formula used along with sample calculations follow.

Formula:

$$\frac{\text{\# 1st alternatives}}{\text{\# responses with multiple alternatives}} = \frac{\text{\# responses with multiple alternatives}}{\text{Total possible number of multiple responses}} \text{ X } \text{\#1st alt.}$$

$$= \text{ reduction from \# 1st alternatives}$$

In the above example:

4/6 = 6/25 X 4 = .96 ==> 4 - 0.96 = 3.04

6/9 = 9/25 X 6 = 2.16 ==> 6 - 2.16 = 3.84